

Low-Density Language Bootstrapping: The Case of Tajiki Persian

Karine Megerdooian and Dan Parvaz

The MITRE Corporation
7515 Colshire Drive, McLean, VA 22102, USA
E-mail: karine@mitre.org, dparvaz@mitre.org

Abstract

Low-density languages raise difficulties for standard approaches to natural language processing that depend on large online corpora. Using Persian as a case study, we propose a novel method for bootstrapping MT capability for a low-density language in the case where it relates to a higher density variant. Tajiki Persian is a low-density language that uses the Cyrillic alphabet, while Iranian Persian (Farsi) is written in an extended version of the Arabic script and has many computational resources available. Despite the orthographic differences, the two languages have literary written forms that are almost identical. The paper describes the development of a comprehensive finite-state transducer that converts Tajik text to Farsi script and runs the resulting transliterated document through an existing Persian-to-English MT system. Due to divergences that arise in mapping the two writing systems and phonological and lexical distinctions, the system uses contextual cues (such as the position of a phoneme in a word) as well as available Farsi resources (such as a morphological analyzer to deal with differences in the affixal structures and a lexicon to disambiguate the analyses) to control the potential combinatorial explosion. The results point to a valuable strategy for the rapid prototyping of MT packages for languages of similar uneven density.

1. Introduction

Low-density languages, for which few online resources exist, raise difficulties for standard approaches to natural language processing, as these statistical systems require a large amount of online training data. Since most of the world languages are considered low-density, new approaches need to be developed to deal with the lack of computational resources these languages present. In this paper, we propose a method for rapid creation of language technology resources for a low-density language by bootstrapping resources from a higher density variety of the language.

There exists a small but significant class of languages which are distributed across two or more orthographies, with each writing system reflecting various social and linguistic factors. One such instance is Persian, which has three distinct main varieties spoken in Iran (sometimes referred to as *Farsi*¹), Afghanistan (also known as *Dari*), and *Tajik* spoken in Tajikistan as well as by the substantial Tajik minority within Afghanistan. Iranian and Afghani Persian are both written in an extended version of the Arabic script, called the Perso-Arabic writing system. Tajiki Persian, however, is usually written using an extended version of the Cyrillic alphabet, and shows the effects of substantial and prolonged contact with Russian, in addition to other central Asian languages. There is a rich set of computational resources for Iranian Persian stretching back at least 30 years – online newspapers, BBS forums, IRC groups, Usenet, gopher, weblogs, etc. – which have been used for developing lexica, morphological analyzers, MT engines, and other tools. The online resources for

Tajiki Persian, however, are extremely scarce and computational systems have not been developed for this lower-density variety of Persian.

The literary written forms of these two varieties of Persian are almost identical; it is therefore possible to take advantage of the relatedness of these languages in order to create resources and build systems for Tajiki Persian with very little effort. This paper presents the strategy developed for Tajik in using the resources available for the higher-density variety language to ramp up an MT solution for lower-density versions of the same language. The system consists of a mapping transducer from Tajik in Cyrillic script to its Perso-Arabic equivalent, a morphological analysis component for Persian combined with lexicon lookup, and a commercial machine translation system from Iranian Persian to English.

2. The Writing Systems of Persian

Correspondences between written Iranian Persian and Tajiki Persian are nontrivial. Tajik orthography is more reflective of Persian phonology than Perso-Arabic script is, while written Iranian Persian takes into account the spellings of borrowed Arabic words and the phonemic structure of Arabic which is quite distinct from Persian. In addition, the two dialects of Persian have developed independently resulting in the distinct pronunciation of certain words, which is also represented in the writing systems.

Iranian Persian or Farsi uses an extended version of the Arabic script; it includes, in addition, the letters for پ /p/, گ /g/, ژ /zh/ and چ /ch/. Although Farsi has maintained the original orthography of Arabic borrowings, the pronunciation of these words have been adapted to Persian which lacks certain phonemes such as interdental and emphatic alveolars. Hence, the three distinct letters س, ص, and ث are all pronounced /s/. One of the main characteristics of Farsi script is the absence of diacritics in

¹ In this paper, the terms *Farsi* and *Iranian Persian* will be used interchangeably; the former is more succinct, and is being used in contrast with *Tajik*.

most written text generally representing the vowels /a/ (as in English *cat*), /e/ (as in *bed*), and /o/, adding to the ambiguity for computational analysis. In addition, the writing system lacks capitalization. Further ambiguities arise due to the fact that in online text, certain morphemes can appear either attached to the stem form or separated from it by an intervening space or control character.

Tajiki Persian is based on the Cyrillic alphabet. It also includes several additional characters that represent Persian sounds not existent in Russian. These are $\text{ҕ} = /h/$, $\text{җ} = /j/$, $\text{қ} = /q/$, $\text{ғ} = /gh/$, $\text{ӯ} = /ö/$, $\text{ӣ} = /i/$. Tajiki text is much less ambiguous than its corresponding Farsi script as all the vowels are generally represented in this writing system and capitalization is used for proper names and at the beginning of sentences. The orthography corresponds more directly to the Persian language pronunciation. For instance, the sounds /s/ and /t/ are represented with the Cyrillic character ‘c’ and ‘r’ respectively, regardless of the original spelling. This has of course created certain homonyms in Tajiki Persian which are differentiated in Farsi orthography. For instance, the two Farsi words of Arabic origin ستر ‘concealment’ and سطر ‘line’, where /t/ is represented with different letters in each instance, are both written as *carp* /*satr*/ in Tajik.

As the two variants of Persian have developed independently for several decades, they have also diverged in pronunciation. Hence, the two distinct pronunciations of *shir* ‘milk’ and *sheyr* ‘lion’ in Tajiki Persian are also represented in the orthography as шир and шеп, respectively, preserving a distinction previously held in Classical Persian, while in Modern Iranian Persian they are both written and pronounced identically as شیر (*shir*). On the other hand, Farsi makes a distinction between *pul* ‘money’ and *pol* ‘bridge’, whereas Tajiki Persian pronounces both as пул (*pol*) (Perry, 2005).



Тоҷикистон бахши умдае аз ниёзҳои аввалияи худ, аз чумла маводди гизои, назирӣ орду равған, маводди сухти, назирӣ нафтӯ бензин ва ҳамчунин порчау либосро аз кишварҳои хоричи ворид мекунад.

تاجیکستان بخش عمده ای از نیازهای اولیه خود، از جمله مواد غذایی نظیر آرد و روغن، مواد سوختی نظیر نفت و بنزین و همچنین پارچه و لباس را از کشورهای خارجی وارد می کند.

Figure 1: Sample Tajiki and Iranian Persian Writing Systems (source: BBC Persian)

Furthermore, Iranian and Tajiki Persian have differing patterns of contact, which leads to different patterns of borrowed words. The choice of orthography makes a difference, as well: whereas Western terms borrowed into Iranian Persian must be reformulated in Perso-Arabic, the use of Cyrillic in Tajik allows for Russian terms (as well as other languages in contact from former Soviet republics, such as Uzbek) to be preserved in the original orthography. For instance, the month October in Iranian Persian is a borrowing from French and is represented as اکتبر /*oktoabr*/ while it is written as in Russian октябрь in Tajiki Persian.

3. Issues in Mapping

As the previous section showed, Iranian Persian and Tajiki Persian, despite being variants of the same language, display a number of divergences in the written representation of words. This section presents some of the issues encountered in a mapping of the Cyrillic script of Tajik into the Arabic-based script used in Iranian Persian.

In certain instances, a basic letter correspondence can help achieve a direct mapping from Tajik into Iranian Persian, as shown in the following examples:

китобҳо	کتابها	<i>ketâbhâ</i>	‘books’
коршиносони	کارشناسان	<i>kârshenâsâne</i>	‘experts of’
Мардум	مردم	<i>mardom</i>	‘people’
вокунише	واکنشی	<i>vâkoneshi</i>	‘a reaction’
Корманди давлати	کارمند دولتی	<i>kârmande dowlati</i>	‘government worker’

Table 1: Direct mapping of Tajiki to Farsi script

However, ambiguities arise at several levels. For instance, the Iranian Persian writing system includes three distinct letters representing the /s/ sound, four characters corresponding to /z/ and two different letters pronounced as /h/, due to the original orthography of the borrowed Arabic words. Hence, a basic mapping to the most common character results in divergences from standard orthography as can be seen by the highlighted characters in the following examples:

ТАҶИК	WRONG MAP	CORRECT FARSI	TRANSLATION
Фурсат	فرست	فرصت	‘opportunity’
Сарвати	سروت	ثروت	‘wealth’
Ҳизби	هلب	حزب	‘political party’
Ҳифз кунад	هفن کند	حفظ کند	‘learns by heart’

Table 2: Ambiguous mapping into Farsi script

Another major divergence comes from the distinct representations of the diacritic vowels – /a/, /e/ and /o/ – in everyday writing. These vowels can be written in many ways in Perso-Arabic script. The /o/ sound, for instance, can be written as the *alef* in اردک /*ordak*/ ‘duck’, the *eyn* in عضو /*ozv*/ ‘member’ or not written at all as in انجمن /*anjoman*/ ‘organization.’ Certain positional cues,

mapped output, the transducer creates all possible results for each input token, which is then disambiguated at the next stage in the process.

```
# Add alef under diacritic at the
# beginning of the word
define initialA [(Aa) <- a || .#. _ ];

# Represent the /a/ sound at the
# end of the word (marked by WD
# tag) as 'he'
define silentH [h <- a %^WD];
```

Figure 2: Contextual rules for mapping Tajiki ‘a’

4.2 Lexicon Look-Up and Disambiguation

For each input token, the resulting transliterated Farsi words undergo morphological analysis and lexicon look-up to determine possible lexical items (Amtrup, 2003). Items for which there is only a single alternative are simply printed as is. In the case of Tajik words for which the FST has produced more than one possible Farsi representation, the alternatives are first stripped of “diacritics” (i.e., “short” vowels and other symbols which are not part of the orthographic skeleton of the word and are not included in current MT systems), then are subject to a number of lookups. First, each alternate form is subjected to a morphological analysis and is looked up in a dictionary. If an analysis is found, then the form is used. If there is no analysis, the word is matched against an unstemmed wordlist culled from various Persian corpora. If a match is found there, that match is used. Finally, if none exists, a number of “rules of thumb” are employed to select a likely alternative based on letter frequencies. Table 3 shows the results of disambiguation when the morphological analyzer/lexicon combination works successfully.

1 alternatives (12 originally) سخنگوی sxngv+Noun+sg+ez [speaker;spokesman;]
1 alternatives (1 originally) بانک bank+Noun+sg [bank;]
1 alternatives (32 originally) تاجیکستان taJykstan+PropN [Tajikistan;]
1 alternatives (12 originally) پرداخته‌است prdaxtn+Verb+ind+perf.past+3sg [pay;attend;]

Figure 3: Disambiguated analyses

If at the end of all these stages more than one match remains, there is no current way of dealing with that eventuality, and so one is chosen at random. At a later stage of this project, a language model will have to be constructed so that a more principled choice may be made.

So if, for instance the Tajik буданд may be rendered as either بودند /budænd/ ‘they were’, or بندند /bædænd/ ‘they are bad’ (after the diacritics are stripped); consequently, the ambiguity might be resolved with a language model which would determine whether a bare verb or a verb phrase with an argument is more likely in that position.

5. Evaluation and Discussion

5.1 Evaluation

Preliminary results are promising, modulo the errors discussed further below. Our current test corpus consists of approximately 500,000 words from articles taken from Radio Ozodi (the Tajik broadcast of Radio Free Europe). As a beginning testbed, this seemed ideal, since the domain largely matches the training corpora of commercial Persian MT systems (in this project, Language Weaver Persian is used), and unlike several other sources, the full range of Tajik diacritics are used; later refinements will have to take into account defective orthography used by many electronic sources. As work progresses, a Golden set will be created which will consist of parallel Tajik/Farsi corpora.

Since at this stage we are not measuring the MT output but rather the mapping accuracy from Tajiki to Farsi scripts, we chose to focus on the disambiguated results of the morphological analyzer for our evaluation. Matching therefore consists of the first two stages of disambiguation as covered in §4.2 above.

At this early stage, a small test set of 6,156 tokens was run through the morphological analyzer and lexical lookup. The results show that the current system is able to achieve 89.8% accuracy in transliterating a document in Tajiki script to its Farsi equivalent. In other words, in the case of 89.8% of input tokens, there was at least one correct transliterated form which was used as input to the MT component. The average token returned with 6.27 alternative spellings. Further analysis on the larger corpus is needed to determine the accurate level of precision and recall for various input documents.

5.2 Discussion of Results

Although the idea of taking advantage of available resources for developing systems or tools for low-resource languages has been exploited before in the literature (cf. Oflazer, Nirenburg and McShane, 2001; Monson et al, 2006; Somers, 2005; Xi and Hwa, 2005), there has been no previous research – as far as we are aware – that implements a transliteration system with the goal of achieving MT capability.

The results of the preliminary evaluation show that the current system is able to achieve close to 90% accuracy for an input corpus that uses the extended version of the Tajiki script. The transliterated document can then be used with the Language Weaver Persian-to-English MT system to create translations of the original Tajiki text. The approach proposed in this paper is only a stopgap measure pending the development of sufficient computational resources to develop more traditional translation software.

Nevertheless, it has been proven effective for rapidly building translation capabilities for a language with scarce resources, in case a related higher-density language with a distinct writing system is available. In the rest of this section we will discuss the main errors encountered, as well as the extensions planned for the system in order to improve the results.

5.2.1. System Errors

A closer examination of the results reveals the main types of errors that our system makes. For example, the lexical issues previously discussed – borrowed words and distinct word-frequency patterns – have yet to be addressed. Distinct Tajik pronunciations of loans and proper names in general are another lexical area which we anticipate addressing in the near term. In addition to proper names, common words also have variant pronunciations (and hence, spellings) as a result of dialectal separation as well as language contact. These variations are not only evident in the Cyrillic script, but were also evident at a time when Tajik was written in Perso-Arabic (Perry 2005). For instance, the Persian سوال is written either as суол /suâl/ or as сувол /suvâl/.

There are also lexical and morphosyntactic issues which need resolving. An instance of lexical issues is the use of words which are peculiar to Tajik which, despite common roots, has been developing separately from Iranian Persian for centuries. One example in the test set illustrates the problem:

Дар катли хабарнигори рус тоҷикҳо гумонбар
мешаванд.

در قتل خبرنگار روس تاجیکها گمانبر می‌شوند.

“The Tajiks were suspected in the murder of the Russian reporter”

The underlined word is unlikely to be found in Iranian texts, and so both lexical lookup and any Farsi MT system are bound to miss the word. Another example illustrates both lexical divergence as well as the use of a grammatical construction not found in Iranian Persian:

Шӯрои мудирияти чунин меҳисобад, ки ба очикистон
лозим меояд маболиғи ро баргардонад.
شورای مدیریتی چنین می‌حسابد که به تاجیکستان لازم می‌آید مبالغه را
برگرداند.

“The supervisory council concluded that Tajikistan would need to remit the amount.”

The first item, *меҳисобад*, is an example both of an idiom “foreign” to Iranian Persian, as well as of a verb which in Persian has fallen out of use, and is replaced by a light verb construction (the old verb *hesâbidan* ‘to figure/reckon’ has been replaced in Iran by *hesâb kardan*, literally ‘to do reckoning’). The fact that Farsi and Tajik have evolved in different directions needs to be addressed in the final system. Also, the construction *лозим меояд* (*lâzem miâyad*, literally to ‘come to need’) is unknown in Iranian Persian, and would have to be replaced by a separate term such as

majbur ast ‘is obligated’. Since the point of this exercise is the translation of low-density languages, the development of an extensive Tajik/Farsi lookup table is unrealistic. Further study will reveal if this approach is beneficial in spite of these challenges.

5.2.2. Extension: Phrasal Boundaries

One of the important distinctions between the Tajiki and Iranian Persian writing systems involves the recognition of phrasal boundaries. Boundary recognition is a significant problem in Iranian Persian which uses the Perso-Arabic script, as there is no capitalization and the main morpheme linking the elements of a noun phrase is pronounced as /e/ and, being a diacritic, is typically not represented in orthography. As expected, this gives rise to very ambiguous results in applications such as MT and entity recognition which involve some level of phrasal parsing. In Tajiki Persian, however, the linking morpheme is represented in text, clearly indicating phrasal boundaries in a sentence. This distinction is illustrated in the following example.

نشست سران کشورهای ساحلی خزر شروع شد

‘The session of the heads of the coastal countries of the Caspian (Sea) began’

The nominal elements in the sentence are linked to each other with the so-called “ezafe” morpheme, which is pronounced as /e/ after consonants and /ye/ after vowels as shown:

Transcription & Gloss:

neshest-e sarân-e keshvarhâ-ye sâheli-e xazar
session-**ez** heads-**ez** countries-**ez** coastal-**ez** Caspian

When a word in the noun phrase does not carry this affix, it marks the phrasal boundary. Hence, in this example, *xæzær* ‘Caspian’ is the end of the NP as shown in the parsed version shown below. However, this /e/ morpheme is typically not written in text, resulting in parsing ambiguity as any of the nouns may present a potential NP boundary for the system.

[نشست سران کشورهای ساحلی خزر] شروع شد]

[NP session-**ez** heads-**ez** countries-**ez** coastal-**ez** Caspian]
[VP beginning became]

Tajiki Persian orthography, on the other hand, explicitly writes the “ezafe” morpheme (*и*) as illustrated below for the same sentence, clearly demarcating the phrasal boundary.

Нишасти сарони кишварҳои соҳили Хазар шуруъ шуд
session-**ez** heads-**ez** countries-**ez** coastal-**ez** Caspian beginning
became

Hence, Tajiki Persian documents provide information on capitalization and boundary recognition which is not available to systems dealing with Iranian Persian text. It is therefore desirable to develop a strategy for transferring such syntactic information which would be relevant in

particular for entity extraction applications of Persian. We will leave this issue for future extensions of the system.

6. Conclusion

This paper presents a methodology for the rapid creation of language technology resources for Tajiki Persian by taking advantage of existing resources and systems developed for the higher-density variety of Iranian Persian. It is expected that in the long term, stopgap systems like the one proposed here will be replaced with fully-developed MT based on the cultivating of resources, parallel corpora, rule development, and so forth. In the meantime, it is hypothesized that this methodology can be used across a variety of unevenly dense languages with distinct scripts, such as Hindustani (Hindi, Urdu), the Turkic languages (Turkish, Azeri, Uzbek, Uighur), and Kurdish (Kurmanji and Sorani).

7. Acknowledgements

This pilot study has been supported by an in-house innovation grant by the MITRE Corporation.

8. References

- Amtrup, J.W. (2003). Morphology in Machine Translation Systems: Efficient Integration of Finite State Transducers and Feature Structure Descriptions. *Machine Translation*, 18(3), pp. 217--238.
- Beesley, K.R., Karttunen, L. (2003). *Finite-State Morphology: Xerox Tools and Techniques*. Palo Alto: CSLI Publications.
- Monson, Ch., Llitjós, A.F., Aranovich, R., Peterson, E. Carbonell, J., Lavie, A. (2006). Building NLP Systems for Two Resource-Scarce Indigenous Languages: Mapudungun and Quechua. In *LREC 2006: Fifth International Conference on Language Resources and Evaluation*. Sydney, Australia, pp. 71--77.
- Oflazer, K., Nirenburg, S., McShane, M. (2001). Bootstrapping Morphological Analyzers by Combining Human Elicitation and Machine Learning. *Computational Linguistics*, 27(1).
- Perry, J.R. (2005). *A Tajik Persian Reference Grammar*. Boston: Brill.
- Somers, H. (2005). Faking it: Synthetic Text-to-Speech Synthesis for Under-Resources Languages – Experimental Design. In *Proceedings of the Australasian Language Technology Workshop 2005*. Sydney, Australia, pp. 71--77.
- Xi, Ch., Hwa, R. (2005). A Backoff Model for Bootstrapping Resources for Non-English Languages. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*. Vancouver, Canada, pp. 851--858.