Orthographic and Morphological Processing for Persian-to-English Statistical Machine Translation

Mohammad Sadegh Rasooli and **Ahmed El Kholy** and **Nizar Habash**
Center for Computational Learning Systems
Columbia University, New York, NY
{rasooli,akholy,habash}@ccls.columbia.edu

Abstract

In statistical machine translation, data sparsity is a challenging problem especially for languages with rich morphology and inconsistent orthography, such as Persian. We show that orthographic preprocessing and morphological segmentation of Persian verbs in particular improves the translation quality of Persian-English by 1.9 BLEU points on a blind test set.

1 Introduction

In the context of statistical machine translation (SMT), the severity of the data sparsity problem, typically a result of limited parallel data, increases for languages with rich morphology such as Arabic, Czech and Turkish. The most common solution, other than increasing the amount of parallel data, is to develop language-specific preprocessing and tokenization schemes that reduce the overall vocabulary and increase the symmetry between source and target languages (Nießen and Ney, 2004; Lee, 2004; Oflazer and Durgar El-Kahlout, 2007; Stymne, 2012; Singh and Habash, 2012; Habash and Sadat, 2012; El Kholy and Habash, 2012). In this paper, we work with Persian, a morphologically rich language with limited parallel data. Furthermore, Persian's standard orthography makes use of a combination of spaces and semi-spaces (zero-width non-joiners), which are often ignored or confused, leading to orthographic inconstancies and added sparsity. We address the orthographic challenge of inconsistent spacing with a supervised learning method which successfully recovers near all spacing errors. We also present a set of experiments for morphological segmentation to help improve Persian-to-English SMT. We show that the combination of orthographic cleanup and morphological segmentation for verbs in particular improves over a simple preprocessing baseline.

2 Related Work

Much work has been done to address data sparsity in SMT employing a variety of methods such as morphological and orthographic processing (Nießen and Ney, 2004; Lee, 2004; Goldwater and McClosky, 2005; Oflazer and Durgar El-Kahlout, 2007; Stymne, 2012; Singh and Habash, 2012; Habash and Sadat, 2012; El Kholy and Habash, 2012), targeting specific out-of-vocabulary phenomena with name transliteration or spelling expansion (Habash, 2008; Hermjakob et al., 2008) or using comparable corpora (Prochasson and Fung, 2011). Our approach falls in the class of orthographic and morphological preprocessing.

Previous research on Persian SMT is rather limited despite some early efforts (Amtrup et al., 2000). A few parallel corpora have been released, such as (Pilevar et al., 2011; Farajian, 2011). We conduct our research on an unreleased Persian-English parallel corpus (El Kholy et al., 2013a; El Kholy et al., 2013b).

In terms of preprocessing efforts, Kathol and Zheng (2008) use unsupervised Persian morpheme segmentation. Other attempts to improve Persian SMT use syntactic reordering (Gupta et al., 2012; Matusov and Köprü, 2010) and rule-based post editing (Mohaghegh et al., 2012). El Kholy et al. (2013a) and El Kholy et al. (2013b) also address resource limitation for Persian-Arabic SMT by pivoting on English.

Our approach is similar to Kathol and Zheng (2008), except that we do not use unsupervised learning methods for segmenting morphemes and we explore POS-specific processing instead of segmenting all words. We make extensive use of available resources for Persian morphology such as the Persian dependency treebank (Rasooli et al., 2013), the Persian verb analyzer tool (Rasooli et al., 2011a), the Persian verb valency lexicon (Rasooli et al., 2011c), and the PerStem Persian segmenter (Jadidinejad et al., 2010).

International Joint Conference on Natural Language Processing, pages 1047–1051,
Nagoya, Japan, 14-18 October 2013.

3 Persian Orthography and Morphology

3.1 Orthography

Persian is written with the Perso-Arabic script. Unlike Arabic, some Persian words have inter-word zero-width non-joiner spaces (or semi-spaces). Many writers incorrectly write the semi-spaces as regular spaces (Shamsfard et al., 2010). This causes data inconsistency and some word-sense ambiguity, e.g., if the word نام‌آشنا *nAm_ĀšnA*[1] 'reputed' (adjective) is written with regular spaces, its meaning becomes 'the familiar name'. While humans may be able to recover, typical natural language processing tools will fail since they expect standard Persian spelling.

3.2 Morphology

Persian has a heavily suffixing affixational morphology with no expression of grammatical gender (Amtrup et al., 2000). We give a brief description of Persian adjectives, nouns and verbs and compare to English.

Adjectives Persian adjectives have a limited inflection space: they may be simple, comparative or superlative. In comparative and superlative forms (except for Arabic loan words), a suffix attaches to the adjective: تر+ +*_tar*[2] '+er' for comparative and ترین+ +*_taryn* '+est' for superlative adjectives. English uses both suffixes ('+er/+est') and multi-word construction with 'more/most', in addition to some irregular cases such as 'good', 'better', and 'best'. As such, it might be hard to define a consistent preprocessing scheme for adjectives in Persian with respect to English.

Nouns Nouns are generally similar to English. For example, like English, a suffix marks plural number: mostly ها+ +*_hA* and sometimes ان+ +*An*. Exceptions include Arabic broken plural loan words. Unlike English, Persian has a suffixing indefinite marker (ی+ +*y*) comparable in meaning to English's 'a' or 'an' indefinite particles. In Persian noun phrases consisting of a noun followed by one or more adjectives, the indefinite suffix attaches to the last adjective.

Verbs A verb in Persian may be inflected in different combinations for tense, mood, aspect, voice and person. There are many interesting phenomena in Persian verbs, e.g. the past tense stem is used with another auxiliary verb to create the future form. When an auxiliary verb is used, prefixes attach to the auxiliary verb instead of the root. The negative marker (+ن *n*+) 'not' and the object pronouns are attached to the verbs, leading to more than 100 verb conjugated forms (Rasooli et al., 2011b). For example, the verb نمی‌خواندمش *nmy_xwAndmš* can be tokenized to *n*+ *my*+ *xwAnd* +*m* +*š* 'I was not reading it' [lit. 'not+ was(continous)+ read(past) +I +it']. Persian is a pro-drop language; almost half of the verbs in the Persian dependency treebank do not have an explicit subject (Rasooli et al., 2013). By comparison, English has a much simpler verbal morphology with explicit subject realization. This suggests that tokenizing Persian verbs may be helpful to Persian-English SMT in that it reduces sparsity and increases symmetry with English.

4 Space Correction

In standard Persian orthography, semi-space characters show inter-word boundaries. Around 8% of all tokens in Persian dependency treebank have semi-spaces (Rasooli et al., 2013). However, in real Persian text, many of these semi-spaces are written as regular space. Although semi-space restoration may actually increase sparsity by creating more compounded forms of words, it is an important step to allow the use of Persian morphological resources that expect their presence.

In order to improve the quality of spacing in Persian texts, we use a language-modeling approach to correct spacing errors. The approach relies on the existence of a lexicon of semi-spaced words. The lexicon provides a mapping model from the regular-spaced versions of the words to their correct semi-spaced version. Starting with a sentence, we identify all sequences of regularly spaced words that can be mapped to semi-spaced versions. An expanded lattice version of the sentence including both forms is then decoded with a language model to select the path with the highest probability.

In terms of resources, we use the Peykare corpus (Bijankhan et al., 2011) and Persian dependency treebank (Rasooli et al., 2013) to create the semi-space lexicon and language model. The training data consists of about 398 thousand sentences and 89 million tokens (12 million types). To construct the lexicon, we extract all words with semi-spaces in the training data. We further extend the lexicon to cover known semi-space inflections for seen words, such as plural suffixes in nouns,

[1] We use the Habash-Soudi-Buckwalter Arabic transliteration (Habash et al., 2007) in the figures with extensions for Persian as suggested by Habash (2010). We show semi-spaces with underscore character.

[2] Suffixes that require a semi-space are marked in the transliteration with an underscore.

superlative and comparative suffixes in adjectives and prefixing continuos markers in verbs. The language model is a trigram model with back-off.

We use the development part of the Persian dependency treebank for tuning the n-gram model. On the test part of the Persian dependency treebank, we replace every semi-space with regular space and try to predict the semi-spaces with our model. The baseline accuracy (of having no semi-spaces) on the test set is 92.2%. Our system's accuracy is 99.43%. The precision, recall and F-score of producing semi-spaces are 93.11%, 99.98% and 96.42%, respectively. The recall of our approach is almost perfect, but the precision is not as good, suggesting that we over assign semi-space. There are two common errors in the results. The first problem is with the hard distinction between adjectives and verbs, e.g., خراب‌شده *xrAb_šdh* 'dilapidated' vs. خراب شده *xrAb šdh* 'has destroyed'. The second problem is with errors in the training data, especially from the Peykare corpus (Bijankhan et al., 2011).[3]

5 Morpheme Segmentation

In this section, we present the two different morphological segmentation methods: PerStem and VerbStem.

PerStem As a baseline method for morphological segmentation, we use the off-the-shelf Persian segmenter, PerStem (Jadidinejad et al., 2010).[4] PerStem is a deterministic tool employing a set of regular expressions and rules for segmenting Persian words. PerStem separates most affixes for all parts-of-speech when applicable. PerStem has been used by other researchers for tokenization purposes (El Kholy et al., 2013a; El Kholy et al., 2013b).

VerbStem As discussed in Section 3, Persian verbs are particularly problematic for Persian-English SMT because of their rich morphology and differences from English. We experiment with targeting Persian verbs for segmentation. To identify which words are verbs, we use a simple maximum likelihood POS tagging model built on the Peykare corpus (Bijankhan et al., 2011). For analysis and segmentation, we use an available Persian verb analyzer tool (Rasooli et al., 2011a)[5] and extend it with a deterministic segmentation algorithm to allow us to generate the needed tokens.[6] For each verb, we segment the negative marker, continuous marker, subject pronoun, object pronoun, participle marker, and prefix marker from the verb stem. We add spaces to the end of prefixes and beginning of suffixes, e.g., نمی‌خواندمش *nmy_xwAndmš* would be segmented into *n my xwAnd m š*.[7] In our segmentation scheme, we do not perform any reordering nor try to address compound verbs in Persian.

Both the POS model and the Persian verb analyzer/segmenter expect the input text to have standard semi-space usage. Thus, we have to apply this step after semi-space correction. Figure 1 presents an example in different representations.

6 MT Evaluation

Experimental Settings We conduct several experiments using different segmentation decisions: **Raw** is original text; **Raw-RS** is Raw text but with regular spaces replacing all semi-spaces; **PerStem** is text processed with PerStem; **Clean-SS** is text with automatically corrected semi-spaces; and **VerbStem** is text processed with the verb segmentation method discussed in the previous section. Figure 1 compares three versions of the same sentence processed in different methods.

We use a Persian-English parallel corpus consisting of about 160 thousand sentences and 3.7 million words for translation model training (El Kholy et al., 2013a; El Kholy et al., 2013b). Word alignment is done using GIZA++ (Och and Ney, 2003). For language modeling, we use the English Gigaword corpus with 5-gram LM implemented with the KenLM toolkit (Heafield, 2011). All experiments are conducted using the Moses phrase-based SMT system (Koehn et al., 2007) with a maximum phrase length of 8. The decoding weight optimization uses a set of 1,000 sentences extracted randomly from the parallel corpus. We use only one English reference for tuning. We report results on a dev set and a blind test set, both with 268 sentences and three English references.

[3] Peykare is not actually written with semi-spaces. However, each word unit (consisting of one or more tokens) is written on one line and it is almost straightforward to standardize the corpus and add the semi-spaces. Unfortunately, some word lines in this corpus have two or more words that should have been written on separate lines, which leads to false examples of inserted semi-spaces, e.g., هنگامی‌که *hngAmy_kh* 'when that' should be written with regular space instead of semi-space.

[4] http://sourceforge.net/projects/perstem/

[5] https://github.com/rasoolims/PersianVerbAnalyzer

[6] We also update the verb list in the Persian verb analyzer using the Persian verb valency lexicon [version 3.0.1] (Rasooli et al., 2011c).

[7] We considered adding plus sign to the end of prefixes and beginning of suffixes, but this representation did worse in SMT experiments.

Input	از فردا نمی‌ترسم چراکه دیروز را دیده‌ام و امروز را دوست دارم
Raw-RS	Az frdA **nmy trsm** crAkh dyrwz rA *dydh Am* w Amrwz rA dwst dArm
	from tomorrow , it would not have seen am yesterday and today i love
PerStem	Az frdA **nmy trsm** crAkh dyrwz rA *dy dh Am* w Amrwz rA dwst dAr m
	from tomorrow , am not seen since yesterday and today i love
VerbStem	Az frdA **n my trs m** crAkh dyrwz rA *dyd h Am* w Amrwz rA dwst dAr m
	from tomorrow , not afraid because i have seen yesterday and today i love
Reference	**i 'm not afraid** of tomorrow because *i have seen yesterday* and i like today

Figure 1: Example output from three systems and one of the references from the dev set. As seen in the bolded and underlined words, the VerbStem system captures linguistic information and produces better translation quality.

Method	**Raw**	**Raw-RS**	**PerStem**	**Clean-SS**	**VerbStem**
BLEU	33.0	33.6	32.6	32.2	**33.7**

Table 1: SMT results on the dev set.

Model	BLEU	METEOR	TER
Raw-RS (Baseline)	31.4	31.2	**60.9**
VerbStem (Best model)	**33.3**	**32.2**	61.1

Table 2: Results from the baseline and the best system on the blind test set.

Results and Discussion The results of SMT experiments on the dev set are shown in Table 1. VerbStem is our best system. Simply replacing all spaces (Raw-RS) does rather well and is plausibly the strongest simplest baseline we can compare to. PerStem and Clean-SS underperform the baseline. Clean-SS is the worst system (as expected since it increases sparsity), but it is necessary as a step for VerbStem. The improvement in VerbStem is possibly the result of reduced sparsity and increased symmetry between English and Persian. Verb segmentation makes a lot of information explicit, such as negation, subject pronoun (especially since Persian as a pro-drop language) and object pronoun.

We apply VerbStem to the blind test set and compare it to Raw-RS. Table 2 shows the blind test results using BLEU-4 (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007) and TER (Snover et al., 2006). VerbStem produces a higher BLEU score improvement over the Raw-RS baseline on the blind test compared to the dev set. This may suggest that our dev set is easier in general. Although our best system does well in Figure 1, the best result still suffers from suboptimal word order. The position of the verb in Persian (as an SOV language) is very problematic when translating to English (an SVO language) especially for long sentences.

7 Conclusion and Future Directions

Our experiments show that segmenting Persian verbs improves translation quality. However, the translation output of all current systems in this paper suffer from word order problems. In the future, we plan to investigate how to improve word order in the translation output using a variety of techniques such as hierarchical phrase-based models (Chiang, 2005; Kathol and Zheng, 2008; Cohn and Haffari, 2013), or models employing parsers to be developed using the Persian dependency treebank (Collins et al., 2005; Elming and Habash, 2009; Carpuat et al., 2010).

Acknowledgments The second author was funded by a research grant from the Science Applications International Corporation (SAIC). We thank Nadi Tomeh for helpful discussions.

References

Jan Willers Amtrup, Hamid Mansouri Rad, Karine Megerdoomian, and Rémi Zajac. 2000. *Persian-English machine translation: An overview of the Shiraz project*. Computing Research Laboratory, New Mexico State University.

Mahmood Bijankhan, Javad Sheykhzadegan, Mohammad Bahrani, and Masood Ghayoomi. 2011. Lessons from building a Persian written corpus: Peykare. *Language Resources and Evaluation*.

Marine Carpuat, Yuval Marton, and Nizar Habash. 2010. Improving Arabic-to-English Statistical Machine Translation by Reordering Post-Verbal Subjects for Alignment. In *ACL'10*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *ACL'05*.

Trevor Cohn and Gholamreza Haffari. 2013. An infinite hierarchical bayesian model of phrasal translation. In *ACL'13*.

Michael Collins, Philipp Koehn, and Ivona Kucerova. 2005. Clause Restructuring for Statistical Machine Translation. In *ACL'05*.

Ahmed El Kholy and Nizar Habash. 2012. Orthographic and morphological processing for English–Arabic statistical machine translation. *Machine Translation*.

Ahmed El Kholy, Nizar Habash, Gregor Leusch, Evgeny Matusov, and Hassan Sawaf. 2013a. Language independent connectivity strength features for phrase pivot statistical machine translation. In *ACL'13*.

Ahmed El Kholy, Nizar Habash, Gregor Leusch, Evgeny Matusov, and Hassan Sawaf. 2013b. Selective combination of pivot and direct statistical machine translation models. In *IJCNLP'13*.

Jakob Elming and Nizar Habash. 2009. Syntactic Reordering for English-Arabic Phrase-Based Machine Translation. In *EACL'09 Workshop on Computational Approaches to Semitic Languages*.

Mohammad Amin Farajian. 2011. Pen: Parallel English-Persian news corpus. In *WORLDCOMP'11*.

Sharon Goldwater and David McClosky. 2005. Improving statistical mt through morphological analysis. In *EMNLP'05*.

Rohit Gupta, Raj Nath Patel, and Ritnesh Shah. 2012. Learning improved reordering models for Urdu, Farsi and Italian using SMT. In *Workshop on Reordering for Statistical Machine Translation*.

Nizar Habash and Fatiha Sadat. 2012. Arabic preprocessing for statistical machine translation. *Challenges for Arabic Machine Translation*.

Nizar Habash, Abdelhadi Soudi, and Tim Buckwalter. 2007. On Arabic Transliteration. In *Arabic Computational Morphology: Knowledge-based and Empirical Methods*. Springer.

Nizar Habash. 2008. Four Techniques for Online Handling of Out-of-Vocabulary Words in Arabic-English Statistical Machine Translation. In *ACL'08*.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Morgan & Claypool Publishers.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *WMT'11*.

Ulf Hermjakob, Kevin Knight, and Hal Daumé III. 2008. Name translation in statistical machine translation: Learning when to transliterate. *ACL'08*.

Amir Hossein Jadidinejad, Fariborz Mahmoudi, and Jon Dehdari. 2010. Evaluation of PerStem: a simple and efficient stemming algorithm for Persian. In *Multilingual Information Access Evaluation*.

Andreas Kathol and Jing Zheng. 2008. Strategies for building a Farsi-English smt system from limited resources. In *INTERSPEECH'08*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL'07*.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *WMT'07*.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *NAACL'04*.

Evgeny Matusov and Selçuk Köprü. 2010. Improving reordering in statistical machine translation from farsi. In *AMTA'10*.

Mahsa Mohaghegh, Abdolhossein Sarrafzadeh, and Mehdi Mohammadi. 2012. GRAFIX: Automated rule-based post editing system to improve English-Persian SMT output. In *COLING'12*.

Sonja Nießen and Hermann Ney. 2004. Statistical Machine Translation with Scarce Resources using Morpho-syntactic Information. *Computational Linguistics*.

Franz Josef Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*.

Kemal Oflazer and Ilknur Durgar El-Kahlout. 2007. Exploring different representational units in English-to-Turkish statistical machine translation. In *WMT'07*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL'02*.

Mohammad Taher Pilevar, Heshaam Faili, and Abdol Hamid Pilevar. 2011. Tep: Tehran english-persian parallel corpus. In *CICLING'11*.

Emmanuel Prochasson and Pascale Fung. 2011. Rare word translation extraction from aligned comparable documents. In *ACL'11*.

Mohammad Sadegh Rasooli, Heshaam Faili, and Behrouz Minaei-Bidgoli. 2011a. Unsupervised identification of Persian compound verbs. In *MICAI'11*.

Mohammad Sadegh Rasooli, Omid Kashefi, and Behrouz Minaei-Bidgoli. 2011b. Effect of adaptive spell checking in Persian. In *NLPKE'11*.

Mohammad Sadegh Rasooli, Amirsaeid Moloodi, Manouchehr Kouhestani, and Behrouz Minaei-Bidgoli. 2011c. A syntactic valency lexicon for Persian verbs: The first steps towards Persian dependency treebank. In *LTC'11*.

Mohammad Sadegh Rasooli, Manouchehr Kouhestani, and Amirsaeid Moloodi. 2013. Development of a Persian syntactic dependency treebank. In *NAACL'13*.

Mehrnoush Shamsfard, Hoda Sadat Jafari, and Mahdi Ilbeygi. 2010. Step-1: A set of fundamental tools for Persian text processing. In *LREC'10*.

Nimesh Singh and Nizar Habash. 2012. Hebrew morphological preprocessing for statistical machine translation. In *EAMT'12*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *AMTA'06*.

Sara Stymne. 2012. *Text Harmonization Strategies for Phrase-Based Statistical Machine Translation*. Ph.D. thesis, Linköping.