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

Mohammad Sadegh Rasooli, Ahmed El Kholy and Nizar Habash

Center for Computational Learning Systems, Columbia University

Introduction

- Our goal is to improve Persian-English SMT via orthographic and morphological processing of Persian.
  - Cleaning Persian orthography to improve morphological segmentation accuracy.
  - Morphological segmentation of Persian words.

Persian Orthography

- Written from right to left with Perso-Arabic script.
- Has some inter-word spaces (zero-width non-joiner or semi-spaces) that are not consistently used by writers.

To the day ...	به روز
Updated	به‌روز
Prosperous	بهروز

Persian Morphology

- Adjectives
  - Suffix 'tar' for comparative (like "er" in English) and 'taryn' for superlative (like "est" in English)

pretty	زیبا
prettier	زیباتر
the prettiest	زیباترین

- Nouns
  - Suffix for plural nouns

book	کتاب
books	کتاب‌ها

- Verbs
  - Inflected in different combinations for tense, mood, aspect, voice and person.

nmy_xwandmš [n+ my+ xwand +m +š]

I was not reading it' [lit. 'not+ was(continous)+ read(past) +I +it'

Space Correction

- This step is mostly needed for our verb morphological analyzer.

Steps

1. Learn a language model from clean data.
2. Get all possible space corrections for a sentence and rank with the language model and return the best spacing.

Space Correction Evaluation

- Train data: BijanKhan corpus and Persian dependency treebank.
- Test data: Persian dependency treebank test data without correct spacing (all semi-spaces are turned into spaces).

Baseline Accuracy	92.20
Correction Accuracy	99.43
Correction Precision	93.11
Correction Recall	99.98
Correction F-Score	96.42

Morpheme Segmentation

We use two morpheme segmentations:

- PerStem: a regular expression matcher that tokenizes all words.
- VerbStem: only tokenizes verbs with high accuracy.
  - We assigned each word the POS with highest probability independent from corpus.
  - Probability of each POS tag is calculated on Bijankhan corpus.

Experimental Results

- Development Results
  - **Raw**: the simple baseline, **Raw-Rs**: After turning semi-spaces into regular spaces, **PerStem**: After using PerStem, **Clean-SS**: After correcting semi-spaces and **VerbStem**: After correcting semi-spaces and segmenting verbs.

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.

- Test Results

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.

- Experimental Setup: Training data: 160K parallel sentences (3.7M words), 1k tuning, 268 dev with 3 English references, 268 blind test with 3 English references, Giza++, KenLM and Moses Decoder for phrase-based SMT.

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.

Conclusion

- Segmenting Persian verbs improves translation quality.
- 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.
- The translation output of all current systems in this paper suffer from word order problems.

Center for Computational Learning Systems, Department of Computer Science, Columbia University

Mail: {rasooli,akholy,habash}@ccls.columbia.edu