

# Supervised Lexical Acquisition for Persian from a Web Corpus

---

**Nick Pendar**

Iowa State University

`pendar@iastate.edu`

**Serge Sharoff**

University of Leeds

`s.sharoff@leeds.ac.uk`

Computational Approaches to Arabic Script-based Languages  
(CAASL)

July 21-23, 2007

Stanford University

# Introduction

- **Problem:** Lack of readily available linguistic resources for Persian
- **Long term goal:** Develop a set of open-domain computational linguistic resources for Persian.
- **So far:**
  - Compiled a Web corpus
  - Developed a stemmer
  - Acquired a lexicon with a supervised approach

# Corpus Compilation

- **Compiled a Web corpus**
  - Random snapshot of the Persian Internet
- **Used Sharoff's (2006) Methodology**
  - Compiled a list of frequent and topic-neutral words
  - Made sure list contains only unambiguously Persian words (as opposed to Arabic)
  - Generated 10,000 trigrams from the list
  - Fed the trigrams to Google as queries
  - Retrieved top ten results & cleaned up

# Corpus Compilation

- **Encoding problem**

- Persian pages use UTF-8, Windows-1256, ISO-8859-6, or MacFarsi encodings
- Encodings often declared incorrectly in source as Latin1
- However 75% of pages retrieved were correct UTF-8
- Retained only UTF-8 pages
- Resulting corpus contains 21,626,774 words

# Preprocessing & Tokenization

- Took out "ـ" (kashideh/tatvil) used to elongate words; e.g, مال مال مال
- Used uniform characters for ک and ی
  - ک showed up as ك or ک
  - ی showed up as ي or ی (U+06CC) or ی (U+0649)
- Spelled out common contractions
  - او است → او است (3sg is)
- Attached orphaned affixes.

# Lexical Profile of Corpus

- 21,626,774 tokens
- 21,267,403 tokens w/ freq.  $\geq 5$
- 318,067 types (~68 avg freq.)
- 238,752 tokens w/ freq.  $< 5$   
(~75% of types)
- **79,315 freq.  $\geq 5$  (~268 avg freq.)  
(98.34% of all tokens)**

# Supervised Acquisition

- Manually build
  - a small seed lexicon
  - a simple lemmatizer
- Run the lemmatizer through the word types
  - If guessed lemma  $l$  for type  $t$  in seed lexicon, or in word types count  $t$  as a form of  $l$ .
  - Else add  $t$  to lexicon.
- Post-edit acquired lexicon

# The Seed Lexicon

- 81 most frequent irregular nouns & their plural forms
- 101 most frequent irregular verbs & their past forms
- 3,123 pronouns, numerals, nouns, regular verbs, and other common words
- Entries assigned one or more parts of speech
  - Tagset with 29 tags; modified version of Penn Treebank Tagset

# The Lemmatizer

- Morphological rules implemented as Python regular expressions.
- If word  $w$  in lexicon: return  $w$  as lemma
- Else:
  - Try morphological rules
  - If guessed lemma  $l$  in lexicon and guessed POS  $p$  equals POS of lemma: return  $l$  as lemma
  - Else:
    - If guesses: return list of guesses; prefer longer lemmas
    - Else: return  $w$  as lemma

# Lemmatizer Rules

- VB →

#indicative

(PRE) (NEG) (HAB) VSTEM (CAUS PAST | PAST\_CAUS) SUBJ  
(OBJ)

#imperative/subjunctive

(PRE) (SUB) VSTEM (CAUS) SUBJ (OBJ)

#past participle

(PRE) (NEG) (HAB) VSTEM (CAUS PAST | PAST\_CAUS)  
PASTP SUBJ

#infinitive

(PRE) (NEG) VSTEM (CAUS PAST | PAST\_CAUS) INF

# Lemmatizer Rules

- NN →

#most common nouns

NSTEM (PL) (POSS | Y) (RA)

#nouns w/ stems ending in ه (h)

NSTEMh (PLh) (POSS | Y) (RA)

#nouns w/ stems ending in و (v/u)

NSTEMv (PLv) (POSS | Y) (RA)

#certain Arabic plurals

ل ل ل ل

# Lemmatizer Rules

---

- JJ → JJSTEM COMP | SUP
- RB → JJSTEM RBSUF
- OD → CD ODSUF

# Lexical Acquisition

- For each word type  $w$ :
  - Lemmatize  $w$
  - If  $\text{lemma}(w)$  in seed lexicon: add  $\text{lemma}(w)$  and  $\text{pos}(w)$  to acquired lexicon
  - If  $\text{lemma}(w)$  observed as word type: add most preferred  $\text{lemma}(w)$  and  $\text{pos}(w)$  to acquired lexicon
  - Else: add  $w$  with no pos to acquired lexicon

# Results

- Acquired lexicon:
  - 43,346 lemmas (~45% reduction in size)
  - 78.3% of lemmas had only one POS tag
  - 17.6% of lemmas had no POS tags
  - 4.0% of lemmas had two POS tags
  - 0.16% of lemmas had three POS tags
  - Avg freq of top 500 lemmas: 11.43
  - Avg freq of bottom half: 1.04

# Results

- POSES assigned to lemmas almost entirely correct.
- Some inaccuracy in assigning lemmas to forms, but post-editing the acquired lexicon far easier than going through the complete word list.
- Lemmas include very few inflected forms.

# Remaining work

---

- Post-editing the acquired lexicon
  - Remove any inflected lemmas
  - Assign POS to lemmas without any POS tags
- Use the cleaned up lexicon for lemmatizer and POS tagger
- Write a complete FST-based morphological analyzer.

# References

---

- Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. In Marco Baroni and Silvia Bernardini (eds.) *WaCky! Working papers on the Web as Corpus*. Bologna: Gedit.