Statistical Natural Language Processing

Language Processing and Language Resources for Persian

Dr. Momtazi

Amirkabir University of Technology

Outline

Character Encoding

- Mixing of Arabic and Persian characters
  (ك vs ک | ي vs ی)
- Sorting Persian characters according to Arabic
  (الف، ب، ت، ث، ج، ح، خ، د، ذ، ر، ز، س، ش، ص، ض، ط، ظ، ع، غ،
  ف، ق، ل، م، ن، هـ، و، پ، چ، ژ، ک، گ، ی)

Word Boundary

- Using space as word boundary
- Multi-token unit problem due to the letters not joined to the next letter (ا، د، ذ، ر، ز،ژ، و)
  - Causes ambiguity in word segmentation (وبا ‘cholera’ vs و با ‘and with’ | مادر ‘mother’ vs ما در ‘we in’)
  - Numbers (بشکه ۱۲۳ vs ۱۲۳بشکه ‘123 barrels’)

Internal Word Boundary

- Using the zero-width non-joiner space (ZWNJS)
- Multi-unit token problem

White-space	ZWNJS	Attached	Transliteration	Translation
می گوید	می‌گوید	میگوید	miguyad	says
هم کلاسی	هم‌کلاسی	همکلاسی	hamkelāsi	classmate
بی نیاز	بی‌نیاز	–	biniyāz	needless
پول ها	پول‌ها	پولها	pulhā	monies
خانه ای	خانه‌ای	–	xāneʔi	a house
بزرگ تر	بزرگ‌تر	بزرگتر	bozorgtar	bigger
بزرگ ترین	بزرگ‌ترین	بزرگترین	bozorgtarin	biggest
بین المللی	بین‌المللی	–	beynolmelali	international
زبان شناسی	زبان‌شناسی	زبانشناسی	zabānšenāsi	linguistics
کتاب سرا	کتاب‌سرا	کتابسرا	ketābsarā	book-house
دانش آموز	دانش‌آموز	–	dānešāmuz	student
علاقه مند	علاقه‌مند	علاقمند	alāqemand	interested
تخم مرغ	تخم‌مرغ	–	toxmemorq	egg
به شیوه	به‌شیوه	بشیوه	bešiveye	like
سنگین وزن	سنگین‌وزن	–	sanginvazn	heavy
در غیر این صورت	درغیراین‌صورت	درغیراینصورت	darqeyreʔinsurat	otherwise
به محض این که	به‌محض‌این‌که	–	bemahzeʔinke	as soon as
صد و بیست و سه هزار	صدوبیست‌وسه‌هزار	–	sadobistosehezār	123000
این کار	این‌کار	اینکار	in kār	this work

Writing Style

- Language varieties
  - Standard: اگر /ʔagar/ ‘if’
  - Super-standard: گر /gar/ ‘if’
  - Sub-standard: اگه /ʔage/ ‘if’

Linguistic Creativity

- Coining a simple spelling for the existing complex word (زنگ زدن /zang zadan/ ‘call’ vs زنگیدن /zangidan/ ‘call’)
- Spelling variation for Arabic words: (حتی vs حتا /hattā/ ‘even’ | حتماً vs حتمن /hatman/ ‘certainly’)

Homographs and Homonyms

- Writing no short vowels: کند
  - /kond/ ‘blunt’
  - /kanad/ ‘picking up’
  - /konad/ ‘doing’
  - /kand/ ‘picked up’
- No capitalization: آذر /āzar/
  - the name of the 9th month in the Persian calendar
  - girl’s name
  - fire

Borrowed Diacritic Characters from Arabic

- Tanvin: جدا /ǰodā/ ‘separate’ vs جداً /ǰeddan/ ‘really’
- Tašdid: بنّا /bannā/ ‘bricklayer’ vs بنا /banā/ ‘building, base’
- Hamze: رئیس /reʔis/ vs رییس /reyis/ ‘boss’
- Short Alef 'ی: 'حتی vs حتی /hattā/ ‘even’

Spelling Variations for Words

- Hamze
- Writing ‘ا’ instead of ‘آ’

  امریکائی /ʔemrikāʔi/
  امریکایی /ʔemrikāyi/
  آمریکائی /āmrikāʔi/
  آمریکایی /āmrikāyi/

- Ezāfe
  - with the intermediary morpheme ‘ی’ /y/ along with a white-space or ZWNJS at the end of the word, such as ‘خانه ی’ or ‘خانه ی’ /xāneye/ ‘the house of’
  - writing ‘ـهٔ’ instead of the intermediary morpheme ‘ی’ /y/ (‘خانهٔ /xāneye/ vs ‘خانه ی” /xāneye/ ‘the house o)
  - without Ezāfe but pronounced: خانه

Foreign Words

- no standard method to write foreign words

  اینترمدییت /intermediyet/
  اینترمدیت /intermediyet/
  اینترمیدیت /intermidiyet/
  اینترمیدییت /intermidiyet/

Contracted Forms

- Phrasal/complex words
- Big challenge in syntactic parsing

  چته /čete/ ‘what is the matter with you?’
  چیه /čiye/ ‘what? | what is it?’
  بچته /bačče/ ’Is it your child?’
  کیستی /kisti/ ‘who are you?’
  کو /ku/ ‘where is it? | that he/she’
  کز /kaz/ ’that from’
  کزو /kazu/ ’that from he/she’

Outline

Available Language Resources for Persian: Text

Name	Designer	Type	Size	Function
Bijankhan	Tehran Uni.	Text	$<$ 2.5 mil. wds	Language modeling
-	Oroumchian et al. [2004]	Text	$<$ 2.5 mil. wds	Data compression
FLDB	IHCS	Text	3 mil. wds	Lexicography
-	Ghayoomi [2004]	Text	$<$ 6.5 mil. wds	Language modeling
-	Darrudi et al. [2004]	Text	$<$ 37 mil. wds	Language modeling
Hamshahri	DBRG	Text	$<$ 37 mil. wds	Information retrieval
PLDB	IHCS	Text	$<$ 56 mil. wds	Lexicography
Peykare	RCISP	Text	$<$ 100 mil. wds	Language modeling
Shiraz Corpus	CRL	Text	3000 Persian-English sentences	Machine translation

Available Language Resources for Persian: Syntax

Name	Designer	Type	Size	Function
Bijankhan Corpus	Bijankhan [2004]	POS taged	$<$ 2.5 mil wds	POS Tagging
PTB	Ghayoomi [2012]	Constituency (HPSG-based)	1,028 sents	Parsing
DPTB	Ghayoomi and Kuhn [2014]	Dependency	1,028 sents	Parsing
UPDT	Seraji et al. [2012]	Dependency	6,000 sents	Parsing
PerDB	Rasooli et al. [2013]	Dependency	29,982 sents	Parsing
Sharif Treebank	Soltanzadeh et al. [2014]	Constituency (Conv. of PerDB)	$<$ 27,000 sents	Parsing

Available Language Resources for Persian: Lexicon|Semantics

Name	Designer	Type	Size	Function
-	Ghayoomi et al. [2015]	Spelling variation	$<$ 30,000 wds	Lexicography, Lan. learning
-	Gerdes and Samvelian [2008]	Complex verbs		Semantics
-	Rasooli et al. [2011]	Word valance		Semantics
-	Ghayoomi [2012]	Named entities	$<$ 2500 nms	Semantics
Farsnet	Shamsfard et al. [2010]	WordNet	$<$ 30,000 ents	Semantics
Persian FrameNet	Nāeblooyi et al. [2015]	FrameNet		Semantics

Available Language Resources for Persian: Discourse

Name	Designer	Type	Size	Function
PCAC	Moosavi and Ghassem-Sani [2009]	Anaphora resolution	2079 prns	NLP, Discourse analysis

Available Language Resources for Persian: Speech

Name	Designer	Type	Size	Function
TFARSDAT	RCISP	Tel. read speech and conversation	about 11 hrs	Speech recognition and caller identification
FARSDAT	RCISP	Mic. speech	25 hrs	Phonetic modeling
The Persian Telephone Conversation Corpus	RCISP	Tel. conversation	about 37 hrs	Speech recognition and language identification
Large FARSDAT	RCISP	Mic. speech	45 hrs	Speech and speaker recognition
The Large Persian Speech Database	RCISP	Conversation speech	$<$ 1000 hrs	Speech recognition
CALLFRIEND Farsi	LDC	Tel. speech	109 calls	Language identification
OGI Multilingual Corpus	OGI	Tel. speech	175 calls	Speech recognition

Reference I

Mahmood Bijankhan. naqše peykarehāye zabāni dar neveštane dasture zabān: mo'arrefiye yek narmafzāre rāyāneyi ["The role of corpora in writing a grammar: Introducing a software"]. *Journal of Linguistics*, 19(2):48–67, 2004.

Ehsan Darrudi, MohammadReza Hejazi, and Farhad Oroumchian. Assessment of a modern Farsi corpus. In *Proceedings of the 2nd Workshop on Information Technology and Its Disciplines*, pages 73–77, Kish Island, Iran, 2004.

Kim Gerdes and Pollet Samvelian. A statistical approach to Persian light verb constructions. In *Proceedings of the 27th International Conference on Lexis and Grammar*, 2008.

Masood Ghayoomi. *pišbiniye vāže dar pardāzeše rāyāneyiye zabāne fārsi ["Word Prediction in Computational Processing of the Persian Language"]*. Master's thesis, Islamic Azad University, Tehran Central Branch, Iran, 2004.

Masood Ghayoomi. Bootstrapping the development of an HPSG-based treebank for Persian. *Linguistic Issues in Language Technology*, 7(1), 2012.

Masood Ghayoomi and Jonas Kuhn. Converting an HPSG-based treebank into its parallel dependency-based treebank. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 802–809, Reykjavik, Iceland, 2014.

Masood Ghayoomi, Saghar Sharifi, and Marziyeh Sanaati. Spelling variation in the Persian language and automatic development of a spelling database from a Web-based corpus. In *Proceedings of the 1st International Conference on Web Research*, University of Science and Culture, Tehran, Iran, 2015.

Nafiseh Sadat Moosavi and Gholamreza Ghassem-Sani. A ranking approach to Persian pronoun resolution. In Alexander Gelbukh, editor, *Advances in Computational Linguistics*, volume 41 of *Research in Computing Science: CICLing '09: Proceedings of the 10th International Conference on Intelligent Text Processing and Computational Linguistics*, pages 169–180. Instituto Politécnico Nacional, M'exico, 2009.

Reference II

Fatemeh Nāeblooyi, Seyyed Mostafa Assi, and Azita Afrashi. šabakeye ma?nāyiye qālebbonyād (framenet) dar zabāne fārsi [frameNet in the Persian language]. *Comparative Linguistics*, (9):261–282, 2015.

Farhad Oroumchian, Ehsan Darrudi, Fattane Taghiyareh, and Neeyaz Angoshtari. Experiments with Persian text compression for web. In *Proceedings of the 13th International World Wide Web Conference*, New York, USA, 2004.

MohammadSadegh Rasooli, Amirsaeid Moloodi, Manouchehr Kouhestani, and Behrouz MinaeiBidgoli. Syntactic valency lexicon for Persian verbs: The first steps towards Persian dependency treebank. In *Proceedings of the 5th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 227–231, Pozna n, Poland, June 2011.

MohammadSadegh Rasooli, Manouchehr Kouhestani, and Amirsaeid Moloodi. Development of a Persian syntactic dependency treebank. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 306–314, Atlanta, Georgia, 2013.

Mojgan Seraji, Beáta Megyesi, and Joakim Nivre. A basic language resource kit for Persian. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 2245–2252, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

Mehrnoush Shamsfard, Hoda Sadat Jafari, and Mahdi Ilbeygi. STeP-1: A set of fundamental tools for Persian text processing. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 859–865, Valletta, Malta, May 19–21 2010. European Language Resources Association (ELRA).

Fatemeh Soltanzadeh, Mohammad Bahrani, and Moharram Eslami. Sharif syntactic treebank: The phrase structure treebank for Persian. In *Proceedings of the 3rd Conference on Computational Linguistics*, Tehran, Iran, 2014.