



Amirkabir University of Technology
(Tehran Polytechnic)

Natural Language Processing

Lecture 19: Persian NLP

Amirkabir University of Technology

Dr Momtazi

Outline

- **Challenges in Persian Text Processing**

Character Encoding

- Mixing of Arabic and Persian characters

(ی vs ي | ک vs ك)

- 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

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
(ما در 'we in' vs مادر 'mother' | و با 'and with' vs وبا 'cholera')
 - Number
(۱۲۳ بشکه vs ۱۲۳ بشکه '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	می گوید	میگوید	miguyad	says
هم کلاسی	هم کلاسی	همکلاسی	hamkelāsi	classmate
بی نیاز	بی نیاز	—	biniyāz	needless
پول ها	پول ها	پولها	pulhā	monies
خانه ای	خانه ای	—	xāneʔi	a house
بزرگ تر	بزرگ تر	بزرگتر	bozorgtar	bigger
بزرگ ترین	بزرگ ترین	بزرگترین	bozorgtarin	biggest
بین المللی	بین المللی	—	beynolmelali	international
زبان شناسی	زبان شناسی	زبانشناسی	zabānšenāsi	linguistics
کتاب سرا	کتاب سرا	کتابسرا	ketābsarā	book-house
دانش آموز	دانش آموز	—	dānešāmuz	student
علاقه مند	علاقه مند	علاقمند	alāqemand	interested
تخم مرغ	تخم مرغ	—	toxmemorq	egg
به شیوه	به شیوه	بشیوه	bešiveye	like
سنگین وزن	سنگین وزن	—	sanginvazn	heavy
در غیر این صورت	در غیر این صورت	درغیراینصورت	darqeyreʔinsurat	otherwise
به محض این که	به محض این که	—	bemahzeʔinke	as soon as
صد و بیست و سه هزار	صد و بیست و سه هزار	—	sadobistosehezār	123000
این کار	این کار	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
(زنگیدن /zangidan/ 'call' vs زدن زنگ /zang zadan/ 'call')
- Spelling variation for Arabic words:
حتا vs حتی /hattā/ 'even'
حتمن vs حتماً /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:

بنا / bannā/ 'bricklayer' vs بنا /banā/ 'building, base'

- Hamze:

رئيس /reʔis/ vs رييس /reyis/ 'boss'

- Short Alef اِ:

حتى /hattā/ 'even' vs حتى

Spelling Variations for Words

- Hamze
- Writing 'ا' instead of 'آ'

امریکائی /ʔemrikāʔi/

امریکایی /ʔemrikāyi/

آمریکائی /āmrikāʔi/

آمریکایی /āmrikāyi/

Spelling Variations for Words

- Ezāfe
 - With the intermediary morpheme 'ی' /y/ along with a white-space or ZWNJS at the end of the word, such as 'خانه‌ی' or 'خانه ی' /xāneye/ 'the house of'
 - Writing ه instead of the intermediary morpheme 'ی' /y/ (خانه /xāneye/ vs خانه ی / xāneye/ 'the house of')
 - Without Ezāfe but pronounced: خانه

Foreign Words

- No standard method to write foreign words
 - اینترمیدیٹ /intermediyet/
 - اینترمڈیت /intermediyet/
 - اینترمیدیٹ /intermidiyet/
 - اینترمیدیٹ /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 /baččate/ 'Is it your child?'
 - کیست /kisti/ 'who are you?'
 - کو /ku/ 'where is it? | that he/she'
 - کز /kaz/ 'that from'
 - کزو /kazu/ 'that from he/she'

Further Reading

- Papers
 - A study of corpus development for Persian
 - M Ghayoomi, S Momtazi, M Bijankhan - International Journal on ALP, 2010
 - Lessons from building a Persian written corpus: Peykare
 - M Bijankhan, J Sheykhzadegan, M Bahrani, M Ghayoomi - Language resources and evaluation, 2011
 - Corpus-based analysis for multi-token units in Persian
 - M Sharifi Atashgah, M Bijankhan, INTERNATIONAL JOURNAL OF INFORMATION AND COMMUNICATION TECHNOLOGY RESEARCH, 2009