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

  (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

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’)

  - Number

    (بشکه ۱۲۳ vs بشکه۱۲۳ ‘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

  (زدن زنگ /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:

  حتیٰ vs حتی /hattā/ ‘even’

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 ـهٔ instead of the intermediary morpheme ‘ی’ /y/ (خانهٔ /xāneye/ vs خانه ی / xāneye/ ‘the house of’)
  - 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 /baččăte/ ’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