بسم الله الرحمن الرحيم

Persian Language Resources Based on Dependency Grammar

Mohammad Sadegh Rasooli
rasooli@cs.columbia.edu

Novermber 2012

Outline

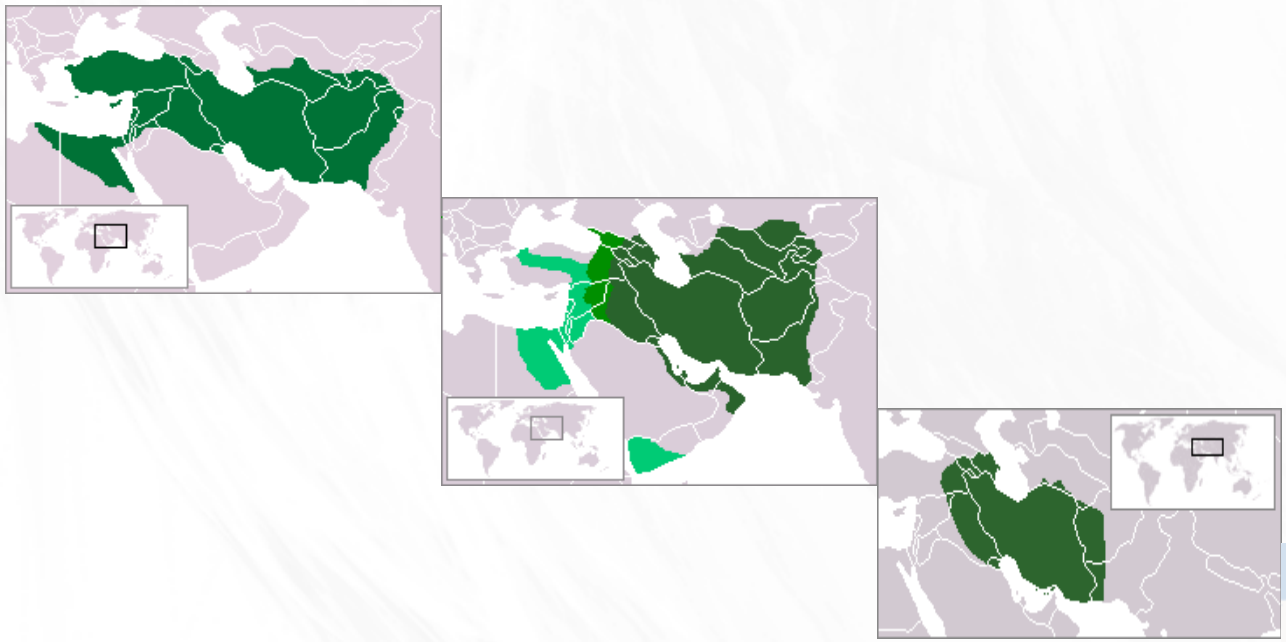
- Iran and Persian Language: An overview
- Challenges in Persian Language Processing
- Persian Resources Based on Dependency Grammar

Iran and Persian Language: An overview

Meaning

- Iran: Land of nobles
- Persia: Land of Persian people
- Persian (Parsi): People from Aryan (Arian) tribe.
- Arya (Aria): Noble (people lived in plateau of Iran).
- Persian language: Language spoken by Persian people.

Iran Map through History

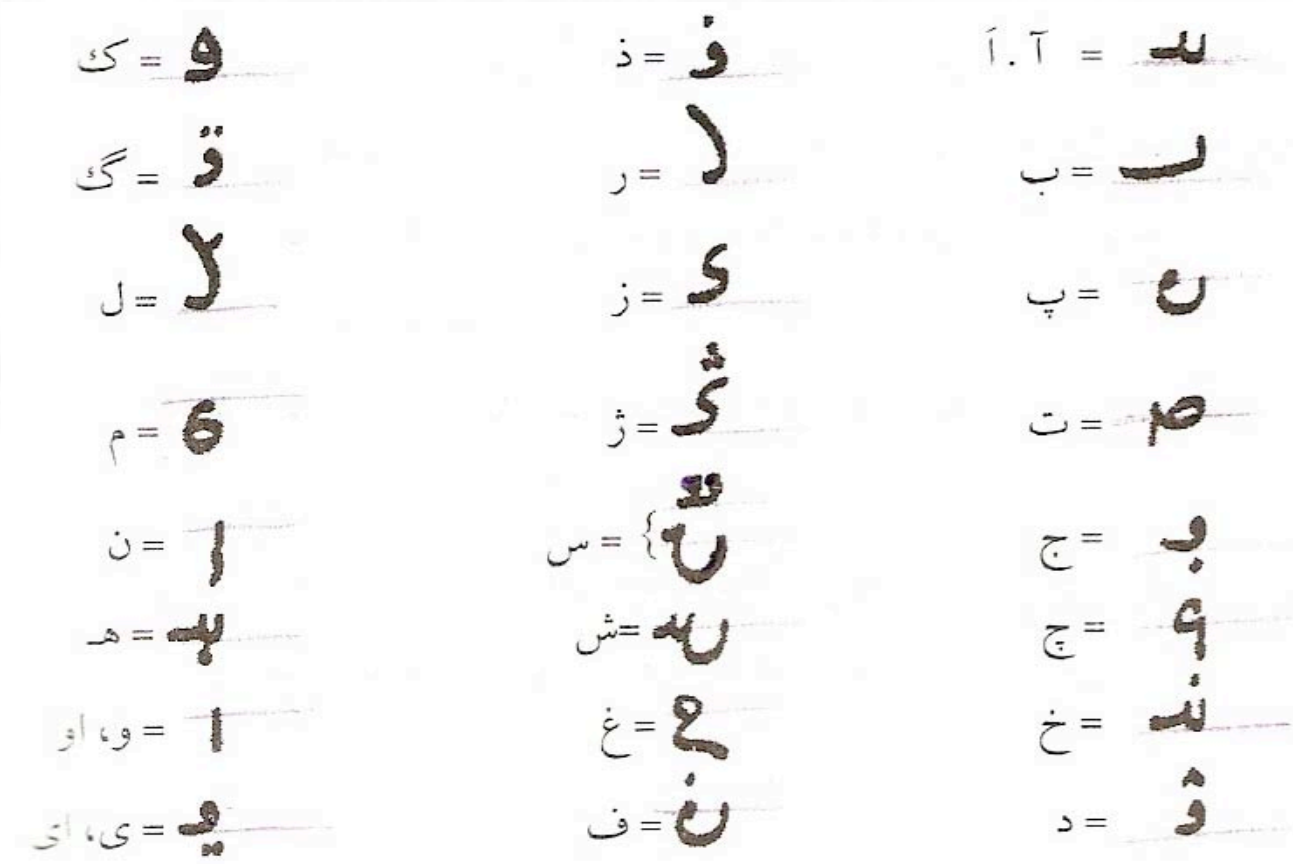


http://en.wikipedia.org/wiki/Greater_Iran

Iran Ethno-religious Distribution

Persian Language in History

- First known as Pahlavi language with Pahlavi script:



Persian Language in History

- Pahlavi script is very similar to Indian scripts.

Persian Language in History

- After Islam, Pahlavi script was replaced by Arabic script with 4 additional characters.

Persian Language in History

- Now, Arabic script is also used in Iran official flag.
- In the middle: الله
- On the horizontal sides: الله اکبر

What is Farsi?

- In standard Arabic there is no “p” sound.
- For 2 centuries, Iran was governed by Arab governors.
- Parsi became Farsi just to be pronounced easier by Arab people.

إذ قال رسول الله: لو كان العلم منوطاً بالثريا لتناوله رجال من **فارس** - بحارالانوار، ١، ١٩٥

Profit Mohammad: Even if knowledge is in the skies, people from **Fars** will gain that knowledge (Behar-al-anvar, 1, 195).

Persian Language

- An Indo-European language
- Written with Arabic script with right-to-left direction.
- Spoken by about 100 million people.
- Now, Persian is the official language in Iran, Afghanistan and Tajikistan.
- In Tajikistan, it is written with Cyrillic script.
  - e.g. نزدیک /naezdik/ наздик

Challenges in Persian Language Processing

Challenges

- Lack of Annotated data
- Colloquial Language
- Orthography
- Morphology
- Syntax

Lack of Annotated Data

- For many open problems in NLP, there is no available Persian corpus.
- Rule based models in Persian did not lead to promising results.

Colloquial Language

- Most of the people use it in their speakings or even their unofficial writings
  - می‌خواهد /miXAhaed/ (he wants)
    - می‌خواد /miXAd/
  - می‌شود /miSaevaed/ (it becomes)
    - می‌شه /miSe/

Orthography

- Diacritics are usually hidden (unless for manual disambiguation)
  - ◌َ /ae/
  - ◌ِ /e/
  - ◌ُ /o/
- سر /s ? r/
  - سُر /sor/: slippy
  - سَر /saer/: head
  - سِر /ser/: secret

Orthography

- Some characters have more than one encoding.
- Affixes are written in multiple shapes (based on the writer style):
  - می‌گویم / می گویم / میگویم
    - “I say”
  - کتابخانه‌ها / کتاب‌خانه‌ها/ کتابخانه ها/ کتاب خانه ها
    - “Libraries”

Orthography

- Semi-space (zero-width non-joiner) is used to attach parts of a unit word, but many people (even experts) do not use it properly.
  - می‌گویم vs. می گویم
    - می /mey/ means “wine” in Persian
    - “I say” vs. “I say wine”
  - خوب‌تر vs. خوب تر
    - تر /taer/ means “wet” is Persian
    - “better” vs. “good wet”

Orthography

- People do not use punctuation between phrases regularly.
- Example (no punctuation, no diacritics):
  - /to/ تو /ketAb/كتاب
    - /ketAb/ /e/ /to/: “Your book”
    - /ketAb/ , /to/: “book, you”

Orthography

- Some Arabic characters have the same pronunciation in Persian:
  - ث س ص /s/
  - ت ط /t/
  - ظ ض ز /z/
- This problem cause ambiguity in speech processing, spell checking, etc.

Morphology

- It is a language with rich morphology.
  - Not as much as Arabic and Turkish
  - تهرانی‌هایشان /tehrAnihAyeSan/
    - “Theirs that are from Tehran”
  - زده‌امشان /zadeaemeSAn/
    - “I have hit them”
  - Arabic words cause irregularity in nouns and verbs

Morphology

- Verbs are the most challenging problem in Persian morphology.
- Types of Persian verbs:
  - Simple
    - Prefix verb
  - Compound verb
    - Prefix compound verb
    - Prepositional phrase verb

Morphology

- Usually, each verb has two lemmas:
  - 1) present and 2) past lemma
    - گفت /goft/ -to speak- (past)
    - گو /gu/ -to speak- (present)
- Verbs (when inflected) can have more than one token:
  - گفت /goft/: “He told”
  - گفته است /gofte aest/: “He has told”
  - گفته خواهد شد /gofte Xahaed Sod/: “It will be told”

Morphology

- Compound verbs:
  - A noun (non-verbal element) with a light verb:
    - صحبت: “speaking”
    - کرد: “to do”
    - صحبت کرد: “to speak”
- Compound verbs can have long distance dependencies (other words can be present between non-verbal element and the light verb)
  - صحبت با تو کردم
    - I spoke with you

Morphology

- Non-verbal elements can also be inflected.
  - صحبت‌های زیادی با تو کردم
  - I spoke with you a lot

Syntax

- Two major problems:
  - Pro-drop
    - Subjects can be omitted easily.
  - Free word order
    - Usually SOV, but others are acceptable.
    - Lots of crossings in syntactic trees.

Persian Resources Based on Dependency Grammar

Motivation

- We developed a spell checker, but there were no syntactic analysis.
- There were no syntactic treebank or lexicons.
- We decided to create
  - A verb valency lexicon (Rasooli et al., 2011)
    - Each verb has what types of complements.
    - More than 4000 verb entries
  - A syntactic treebank

Syntactic Representation

- There were two main options:
  - Generative Grammar
    - e.g. Penn Treebank (Marcus et al., 1993)
  - Dependency Grammar
    - e.g. Prague Dependency Treebank (Böhmová et al., 2003)
- We selected dependency grammar
  - WHY?

Syntactic Representation

- Both of the representations have the ability to show the language structure.
- Dependency grammar is a better choice for free-word order languages (Oflazor et al., 2003).
  - In most of the languages, there are dependency treebanks.
    - There are at least 30 languages with available dependency treebanks (Zeman et al., 2012).
- Dependency representation is more similar to the human understanding of language (Kübler et al., 2009).

Treebanking

- Phase 1: Research and annotation manual documentation.
- Phase 2: Annotating 5000 independent sentences from official online Persian news and websites.
  - With bootstrapping approach.

Treebanking

- Problems with Phase 2:
  - Most of texts are from news texts.
  - From ~5000 verbs in the valency lexicon, only 20% of them were seen at least once.
    - It is impossible to capture all verbs in news texts.
      - We also needed this data for educational needs.

Treebanking

- Phase 3:
  - Collecting sample sentences with unseen verbs from web.
    - About 5-9 random sentences for each verb.
- Phase 4:
  - Collecting common errors in the treebank and revise them manually.

Statistics

- 44 dependency relations
- 17 coarse-grained POS tags
  - Lemmas and some morphosyntactic features have been annotated manually.
- 29,982 sentences
  - 80% train, 10% dev., 10% test
  - Average length: 16.61
- 498,081 words
  - 37,618 unique words
  - 22,064 unique lemmas

Statistics (Verbs)

- 60,579 verbs
- 4,782 unique lemmas
- Average frequency: 12.67

Statistics (Annotator Agreement)

- Sentences were annotated once (plus one more time revision).
- 5% of the sentences were randomly selected to be annotated twice by two different annotators:
  - Labeled dependency relation: 95.32%
  - Dependency relation: 97.06%
  - POS tags: 98.93%

Statistics (Revisions)

- After correcting common errors, the following changes have been made:
  - Labeled dependency relation: 04.91%
  - Dependency relation: 06.29%
  - POS tags: 04.23%

Parsing Accuracy

- Two reported accuracies on version 0.1 (not 1.0):
  - (Zeman et al., 2012)
    - 1.77% nonprojectivity (arc crossing) in version 0.1
    - 86.84% unlabaled attachment score with Malt Parser stack-lazy algorithm (Nivre et al., 2007).
  - (Khallash, 2012)
    - Ensemble model (Malt and MST parser)
    - Best labeled accuracy: 85.06

Conclusions

- It is very hard to have 2 annotators agree one the same syntactic tree.
  - We had 14 annotators.
- It is very hard to have a unique writing style.
  - We tried to trade off between a standard style and keeping the source text writing style.

Dadegan Research Group

Mohammad Sadegh Rasooli

Manouchehr Kouhestani

Amirsaeid Moloodi

Farzaneh Bakhtiary

Parinaz Dadras

Maryam Faal-Hamedanchi

Saeedeh Ghadrdoost-Nakhchi

Mostafa Mahdavi

Azadeh Mirzaei

Yasser Souri

Sahar Oulapoor

Neda Poormorteza-Khameneh

Morteza Rezaei-Sharifabadi

Sude Resalatpoo

Akram Shafie

Salimeh Zamani

Seyed Mahdi Hoseini

Alireza Noorian

Obtain Data

http://dadegan.ir/en

با سپاس از توجه شما

References

- Böhmová, Alena, Jan Hajic, Eva Hajicová, and Barbora Hladká. "The prague dependency treebank: Three-level annotation scenario." Treebanks: Building and Using Parsed Corpora 20 (2003).
- Khallash, Mojtaba, "A mechanism for exploring of the effect of different morphologic and morphosyntactic features on Persian dependency parsing", Master Thesis, Iran University of Science and Technology, 2012.
- Kübler, Sandra, Ryan McDonald, and Joakim Nivre. "Dependency parsing." Synthesis Lectures on Human Language Technologies 1, no. 1 (2009): 1-127.
- Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. "Building a large annotated corpus of English: The Penn Treebank." Computational linguistics 19, no. 2 (1993): 313-330.
- Nivre, Joakim, Johan Hall, and Jens Nilsson. "Maltparser: A data-driven parser-generator for dependency parsing." In Proceedings of LREC, vol. 6, pp. 2216-2219. 2006.

References

- Oflazer, Kemal, Bilge Say, Dilek Zeynep Hakkani-Tür, and Gökhan Tür. "Building a Turkish treebank." Treebanks (2003): 261-277.
- Rasooli, Mohammad Sadegh, Amirsaeid Moloodi, Manouchehr Kouhestani, and Behrouz Minaei-Bidgoli. "A syntactic valency lexicon for Persian verbs: The first steps towards Persian dependency treebank." In 5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics, pp. 227-231. 2011.
- Zeman, Daniel, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. "Hamledt: To parse or not to parse." In Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC'12), Istanbul, Turkey. 2012.