Persian Language, is Stemming Efficient?

Ljiljana Dolamic, Jacques Savoy

Computer Science Department

University of Neuchatel, Switzerland
www.unine.ch/info/clef/

Outline

- □ Introduction
- □ Persian Language
- □ Indexing strategies
- □ IR models
- □ Evaluation
- □ Conclusion

Introduction

- Persian language -50 million native speakers
- Test-collection
  - University of Teheran, CLEF 2008
  - Hamshahri (1996-2002)
  - 166,477 documents
  - 100 topics
    - 501-550 – training topics for CLEF 2008
    - 551-600 – test topics for CLEF 2008
  - Relevance judgement

Introduction

□ <topic lang="**fa**">

<identifier>**10.2452/584-AH**</identifier>

<title>**تحولات قيمت نفت**</title>

<description> **تحولات قيمت نفت در بازار ايران و جهان چگونه است؟**</description>

<narrative> **مي خواهم آمار تحولات قيمت نفت در بازارهاي جهاني و تاثير آن روي اقتصاد و تورم چيست؟**</narrative>

</topic>

Outline

- □ Introduction
- □ Persian Language
- □ Indexing strategies
- □ Models
- □ Evaluation
- □ Conclusion

Persian Language

- □ Indo-European language
  - ■ Distinctly related to European languages
- □ Written in Arabic script (28 letters)
  - ■ Additional four letters
    - □ گ ژ چ پ
      - ■ پدر – father (pedar)
      - ■ بچّه – child (bačče)
      - ■ ژاله – dew (žāle)
      - ■ مرگ – death (marg)

Persian language – writing system

- □ Written from right to left
- □ Four different forms to a letter
  - ■ ت – teh
    - □ توت – unconnected and initial
    - □ رست – final
    - □ چتر – medial
- □ Letters ا, د, ذ, ر, ز, ژ and و do not unite with the letter following

Persian language - vowels

- All letters are consonants
  - ی و ا – used to express vowel sounds ‘دو’ (two)
- Short vowels
  - Vowel signs ‘دَر’, ‘پُشت’, ‘دِل’
- Long vowels
  - Vowel sign + ی و ا - ‘کَار’
- Other signs
  - ٓ, ٔ, ْ, ّ

Persian language

- Definite article
  - کتاب - “book” or “the book”
  - آب بیار (bring water) - آب را بیار (bring the water)
  - مردی که - the man who
  - پسره - son (in question)

Persian Language

- Indefinite article
  - کتاب (book) – کتابی (a book)
  - یک کتاب –a book
  - یک پادشاهی –a king
  - خانهٔ ی –a house
  - نامهٔ –a letter
  - مردهائ ی – some men

Persian Language

- No grammatical gender
  - مرد – man, زن – woman
  - شیرِ نر – lion, شیرِ ماده – lioness
  - Arabic words
    - Danser: رقاص (male) - رقاصه (female)

Persian Language

- Declination
  - Accusative: زنرا (the woman), مرد را (the man)
  - Genitive: خانهٔ مرد (the man’s house)
    پسرِ مرد (the man’s son)

- Prepositions for expressing relations
  - از مرد (from the man)

Persian Language

- □ Plural endings:
  - ■ ان (animate), ها (inanimate)
    - □ پدران, آقایان, بنده → بندگان
    - □ گلها, خانهها, مردها
  - ■ Arabic plurals:
    - □ Sound plurals: ین (masculine) , ات (feminine)
    - □ Broken plurals: قلب → قلوب
- □ Adjectives
  - ■ Substantives: plural endings بزرگان (the great)

Persian language - problems

- Vowel signs omitting
  - مرده (مُرده) – dead vs. مرد (مَرد) – man
- ZWNJ
  - خانه‌ها and not خانه ها or خانهها
- Arabic vs. Persian letters
  - ي (\u064A), ى (\u0649) vs. ی (\06CC)
  - ﻱ, ﻲ, ﻴ, ﻳ vs. ﯼ, ﯽ, ﯿ, ﯾ

- □ Introduction
- □ Persian Language
- □ Indexing strategies
- □ Models
- □ Evaluation
- □ Conclusion

Indexing strategies - stemming

- “none”
- Light stemmer - www.unine.ch/info/clef/
- perstem
  - Jon Dehdari
  - http://sourceforge.net/projects/perstem/
  - *“If these regular expressions are readable to you, you need to check-in to a psychiatric ward!”*
- 5-gram

Indexing strategies - stoplist

- Available at : www.unine.ch/info/clef/
  - 881 term
  - Non content baring terms
  - Suffixes and prefixes
    - Ex.: plural suffixes ان or ها

Outline

- □ Introduction
- □ Persian Language
- □ Indexing strategies
- □ IR models
- □ Evaluation
- □ Conclusion

IR models

- □ Vector-space
  - *Lnu-ltc*
  - *tf-idf*
- □ Probabilistic
  - Okapi
  - DFR-*PL2*
  - Language model LM

Outline

- □ Introduction
- □ Persian Language
- □ Indexing strategies
- □ IR models
- □ Evaluation
- □ Conclusion

Evaluation

T query	none	light	perstem	5-gram
tf idf	0.2966	0.2625*	0.2799*	0.2814
Lnu-ltc	0.4533	0.4277*	0.4369	0.4107*
Okapi	0.4811	0.4535*	0.4610*	0.4511
DFR-PL2	**0.4939**	**0.4693***	**0.4750***	**0.4423***
LM	0.4348	0.4000*	0.4113*	0.3892*
over “none”		-6.79%	-4.43%	-8.57%

Evaluation

- □ Topic #525 “Seasonal Diseases”
  - فصلي (seasonal) → فصل (season, term, article)
  - فصل كم كاري - layoff
- □ Topic #522 “Teheran metro project”
  - متروي (metro)
  - مِتْروي
  - مِتِر (meter)

Evaluation

T query	none		perstem	
	stoplist	no stoplist	stoplist	no stoplist
tf idf	0.2966	0.2449*	0.2799	0.2341*
Lnu-ltc	0.4533	0.4519	0.4369	0.4363
Okapi	0.4811	0.4806	0.4610	0.4591
DFR-PL2	**0.4939**	**0.4888**	**0.4750**	**0.4747**
LM	0.4348	0.4186*	0.4113	0.4034
over “none”		2.74%		3.47%

Evaluation

- Avdl
  - 2.79(stop list) vs. 3.05 (no stop list)
- TD queries
  - 5.94 (stop list) vs. 9.62 (no stop list)
  - TD queries, light stemmer

TD query	*tf idf*	Okapi	DFR-PL2	Over no stop list
stoplist	0.2744	0.4559	0.4785	11.55%
no stoplist	0.2176	0.4110	0.4576	

Outline

- □ Introduction
- □ Persian Language
- □ Indexing strategies
- □ IR models
- □ Evaluation
- □ Conclusion

Conclusion

- Main aspects of Persian morphology
- Indexing strategies
  - Stemming hurts retrieval effectiveness
  - Effect of the stop list removal depends on query size
- IR models
  - *Divergence from randomness*

Persian Language, is Stemming Efficient?

THANK YOU!

Ljiljana Dolamic, Jacques Savoy

Computer Science Department

University of Neuchatel, Switzerland

www.unine.ch/info/clef/