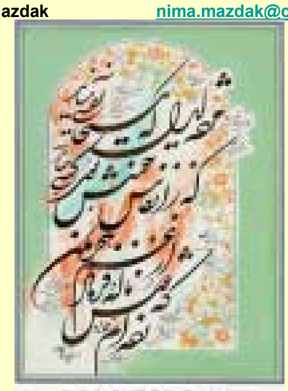
FarsiSum
A Persian text summarizer خلاصه نویس متون فارسی
Master Thesis 20 credits
Nima Mazdak nima.mazdak@comhem.se



Purpose

To implement a Persian text summarizer:

- using the techniques and algorithms developed in the SweSum project
- handling of text containing Unicode characters SweSum supports only ASCII
- adding some new modules , Stop List

To evaluate the summarizer

Background

There are two major types of text summary: ***abstract*** and ***extract***.

Abstract Summarization

The summarized text is an interpretation of the original text.
The process of producing it involves rewriting the original text in a shorter version by replacing wordy concepts with shorter ones.
Example:

He ate banana, orange and pear" can be summarized as

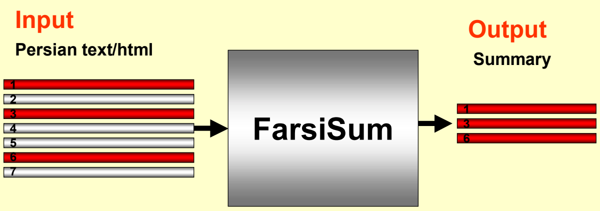
He ate fruit

Not easy to implement

Extract Summarization

Extract Summarization

The summarized text is extracted from the original text on a statistical basis or by using heuristic methods or a combination of both.



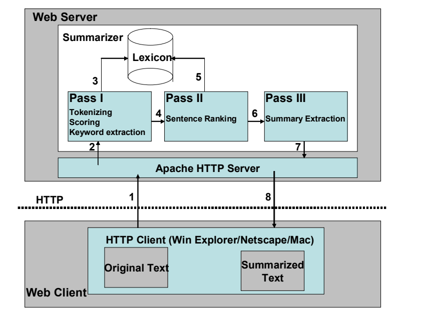
Input
Persian text/html

FarsiSum

Output
Summary

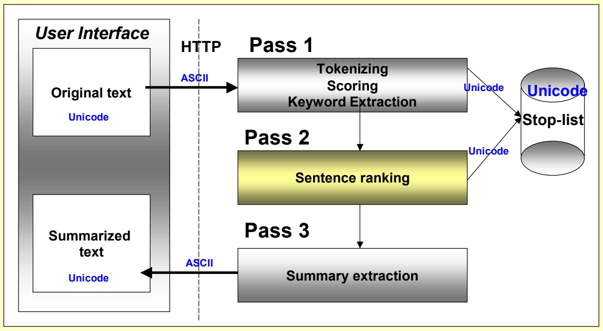
A Client/Server web-based application

SweSum
Archtecture



Alphabet	Roman/Persian
Encoding	ASCII/Unicode
Data	Lexicon/Stop List

FarsiSum
Archtecture



Alphabet

- Right to left
- 4 forms initial, medial, final, isolated
- Letters in a word are connected
- Last character in the word marks the end

آب	ĀB	water
باد	bāD	wind
پیر	pīR	old

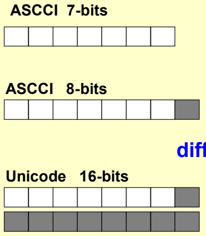
Initial	Medial	Final	Isolated	
گـ	ـگـ	ـگ	گ	G

Short vowels: a e u

Long vowels: ā ī ū
↓ ↓ ↓
آ y v

dar	→	dr	(door)
gul	→	gl	(flower)
nima	→	nyma	
Mazdak	→	mzdk	

Unicode



1111111 = 127

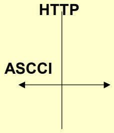
] } Å å
[{ Ä ä
\ | Ö ö

11111111 = 255

different encodings &fonts for Persian

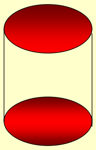
1111111111111111 = 65535

Persian character → a unic code



Character	Unicode	UTF-8
B ب	0628	&#1576;
D د	062F	&#1583;

The Stop List



HTML file (UTF-8 encoding) containing about 200 high-frequency Persian words including:

the most common
verbs, pronouns, adverbs,
conjunctions, prepositions and articles.

Words not included in the stop-list are

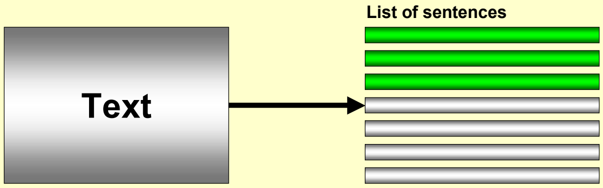
Content Words: nouns or adjectives

The Stop List

The stop-list has been successively built during the implementation phase by running FarsiSum in order to find the most common words in Persian.

Pronoun	من آقا آقای آقایان آن آنان آنکه آنها او این ایشان اینکه این برخي تو خود خودم خودمان خویش شما ما
Conjunction& quantifier	آیا اما اگر البته اول اولین ای چند چه دوم که می و ولی ها هم هر یا یعنی
Adverb	آنجا اکنون امروز اینجا بسیار بسیاری بطور بیش تمامی جا چنان چنین حقیقتا علیرغم فقط همان هیچ هنوز
Preposition	از با بدون بجز بر برای به بی پس پیش تا توی توسط جز داخل در درباره درین را روی سوی علیه غیر کنار میان
Verb	افزود است باشد باشید باشیم بدهید بکنید بگذاریم بگوییم بماند بود بودند بوده خواهد خواهند داد شود کرد کردم کردن | دادم دادند داده دارد دارند داریم داشت داشته دانند رسیده شد شدند شده رسد می‌داند می‌تواند باشد می‌کردند کرده کند کنید کنیم گردند گرفت گرفته گفت گفتم گفتند گفته می خواهیم نماید نموده نیست نیستند کنم ندارد ندارم ندارند نداشته نمی‌کند می‌شود می‌کنند می‌کنم می هست هستیم هستند

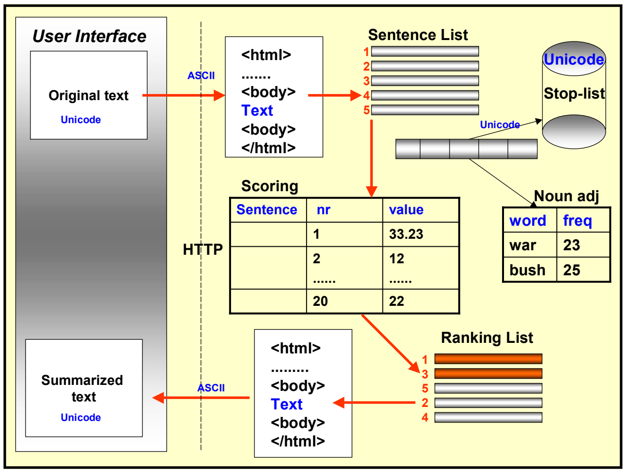
Pass I, II, III



- Tokenize ! . ؟ ؛ ،
- Scoring of sentences
- Sort
- Extract

Scoring

- *First line*	high score. (Default value '1000')
- Position	most important first line followed by other lines *Position score* = (1/line nr)*10.
- Numerical	dates 2004-01-01
- Bold	<B> Bold text in the HTML </B> (100) **Bold text in the HTML**
- Keywords	the most frequent words in the text
- User keywords	



Sentence	nr	value
	1	33.23
	2	12 ......
	20	22

word	freq
war	23
bush	25

Tokenizer

- Converts ASCII/UTF8
- Removes all new line characters “\n”
- Marks all abbreviations (SweSum)
  ex: <! ABBRV>sv. </! ABBRV>.
- Invokes Pronominal Resolution (SweSum)
  John kissed Lisa. He has been in love with her
- Finds the sentence/word boundaries by searching for periods, exclamations, question marks and <BR>
  . , ! ? < > : spaces tabs

Tokenizer

The output of the tokenizer

Sentence	Line Nr
<html>	1
<title> War against Iraq </title>	2
<body>	3
American-led forces will stay in Iraq no longer than necessary.	4
....	....
</body>	n-1
</html>	n

Keyword Extraction

Word	Frequency
American-led	10
Force	5
Iraq	26
Baghdad	13
....	...
War	20

Scoring

Sentence	Line Nr	Value
<html>	1	Not text
<title> War against Iraq </title>	2	Not text
<body>	3	Not text
American-led forces will stay in Iraq no longer than necessary.	4	Text
....	....	
</body>	n-1	Not text
</html>	n	Not text

Scoring

- *First line* high score. (Default value ‘1000’)
- Position most important first line followed by other lines
  Position score = (1/line nr)*10.
- Numerical
  dates 2004-01-01
- Bold <B> Bold text in the HTML </B> (100)
  Bold text in the HTML
- Keywords the most frequent words in the text
- User keywords

Scoring

Example:

Word score = (word frequency) * (a keyword constant

Sentence Score = ∑ *word score* (for all W in sentence)

Example:

American-led forces will stay in Iraq no longer than necessary

Word score (American-led) = 10 * 0.333 => 3.33

Word score (force) = 5 * 0.333 => 1.665

Word score (Iraq) = 26 * 0.333 => 8.658

Sentence Score = 3.33 + 1.665 + 8.658 => **13.653**

Scoring

Example:

Average sentence length **(ASL) =** ***Word-count / Line-count***

Sentence score **=**

(ASL * ***Sentence Score*****)/ (nr of words in the current sentence)**

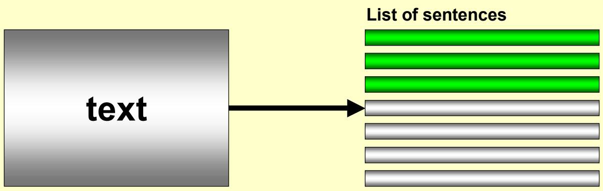
Ex:

ASL = *Word-count / Line-count*= 40/5=> 8

Nr of words in the current sentence = 10

Sentence score **= 13.653**

Sentence Score **= (8*13.653)/10 = 10.9224**



- **Tokenize ! . ؟ ؛ ،**
- **Scoring of sentences**
- **Sort**
- **Extract**

Evaluation of FarsiSum

Individuals have very different ideas on what a good summary should contain

40-70% agreement (Hassel)

Seven native speakers

subjectively compare three different summaries generated by three different methods.

•*Stop-list enabled:* access to the stop-list

•*Stop-list disabled*: No access to the stop-list.

•*The generic mode in SweSum*:

- **the Persian comma, semi colon and question mark are not recognized as sentence/word boundaries.**
- **no access to the stop-list**
- **Final verbs are not removed**

Evaluation of FarsiSum

questions:

•Which summary was the best one?

•Given a scale of 1-5 (1 for the lowest), what score would you assign to each summary?

•Which summary was the most coherent one?

•Which summary preserved the most important information?

Method

	Text	M1	M2	M3
Text	T1	57,1%	14,3%	28,6%
	T2	57,1%	0%	42,9%
	T3	42,8%	28,6%	28,6%
	Average	*52,3%*	*14,3%*	*33,4%*

Table 21: The best method

Evaluation of FarsiSum

Text	M1	M2	M3
T1	39,1%	33,3%	27,6%
T2	37,5%	27,8%	34,7%
T3	35,8%	29,9%	34,3%
Average	*37,5%*	*30.3%*	*32,2%*

Table 22: The best method (score of 1-5)

Evaluation of FarsiSum

Text	M1	M2	M3
T1	50,0%	30,0%	20,0%
T2	36,4%	27,2%	36,4%
T3	36,4%	36,4%	27,2%
Average	*40,1%*	*31,2%*	*27,9%*

Table 23: Cohesion

Evaluation of FarsiSum

Text	M1	M2	M3
T1	44,4%	33,3%	22,3%
T2	77,8%	11,1%	11,1%
T3	41,7%	25%	33,3%
Average	*54.6%*	*23,2%*	*22,2%*

Table 24: Important information preserved

Ambiguity in persian morphology

- Word/Phrase boundary
- Morphology
- Possessive construction
- Light Verb construction

Ambiguity in morphology

a e u

ā ī ū

↓ ↓ ↓

آ y v

dar → dr (door)

gul → gl (flower)

کرم *krm*					
kerm	*karam*	*karam*	*kerem*	*krom*	*karm*
worm	generosity	name	cream	chrome	vine

Different meaning of words

Keyword (word freq.)

Phrase Ambiguity

Initial	Medial	Final	Isolated	
ﮔ	ﮕ	ﮓ	ﮒ	G

Mellanslag

jaGheteRnimA jaG heteR nimA

Problem

- Ord/phrase ambiguity
- Fri/bunden morpheme
- Compound word

Keywords in text summarizer

Solution:

modify tokenizer

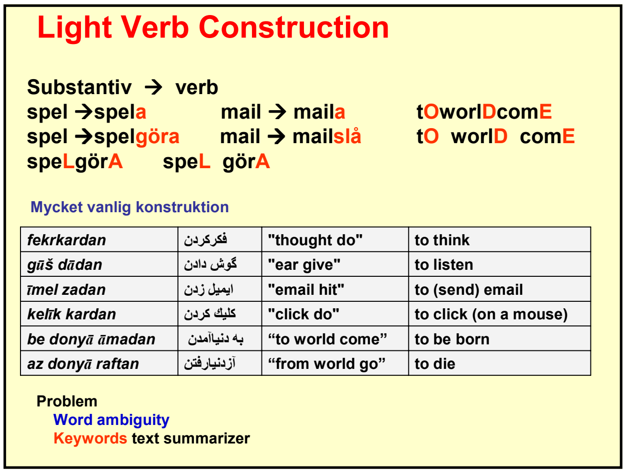
access to lexicon, parser

Fri/Bound morphemes

Free morpheme with space	Free morpheme without space	bound
مى روم	مى‌روم	میروم
mĪ rvm	mĪrvm	mīrvm
mī ravam (I go)		

example (affix *mī*)

- As free morpheme *mĪ* with space between *mĪ* and *ravam*
- As free morpheme *mĪ* without space
- As bound morpheme



Light Verb Construction

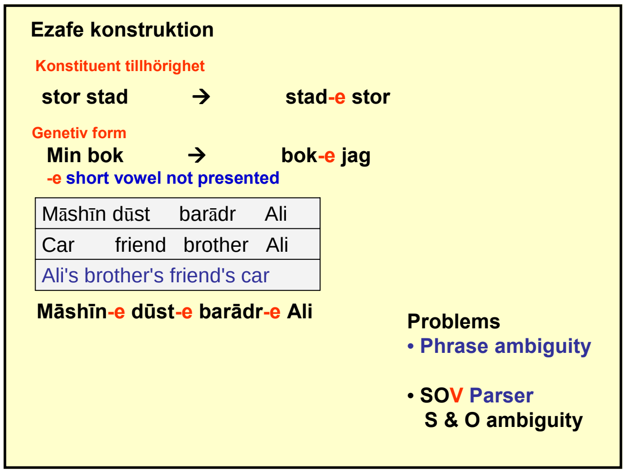
Substantiv → verb

spel →spela mail → maila tOworlDcomE
spel →spelgöra mail → mailslå tO worlD comE
speLgörA speL görA

Mycket vanlig konstruktion

fekrkardan	فكركردن	"thought do"	to think
gūš dādan	گوش دادن	"ear give"	to listen
īmel zadan	ايميل زدن	"email hit"	to (send) email
kelīk kardan	كليك كردن	"click do"	to click (on a mouse)
be donyā āmadan	به دنياآمدن	“to world come”	to be born
az donyā raftan	آزدنيارفتن	“from world go”	to die

Problem
Word ambiguity
Keywords text summarizer



Ezafe konstruktion

Konstituent tillhörighet

stor stad → stad-e stor

Genetiv form

Min bok → bok-e jag

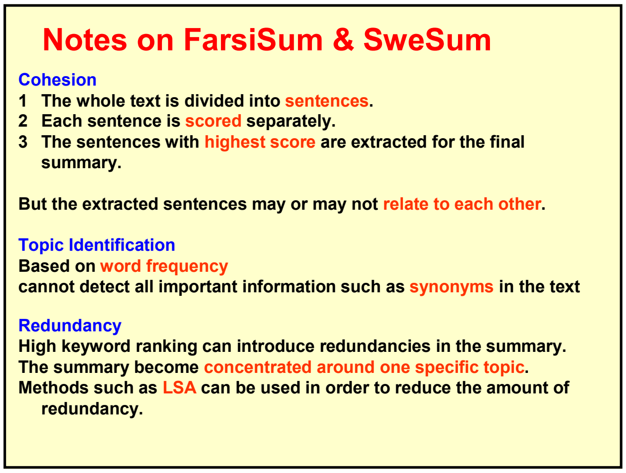
-e short vowel not presented

Māshīn	dūst	barādr	Ali
Car	friend	brother	Ali
Ali's brother's friend's car			

Māshīn-e dūst-e barādr-e Ali

Problems
- Phrase ambiguity

- SOV Parser
  S & O ambiguity



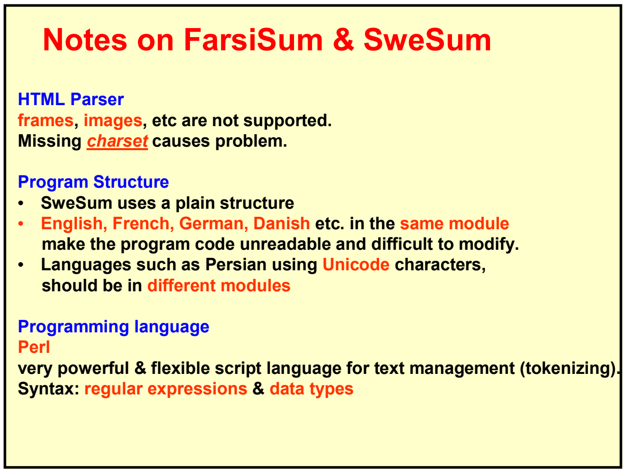
Notes on FarsiSum & SweSum

Cohesion
1 The whole text is divided into sentences.
2 Each sentence is scored separately.
3 The sentences with highest score are extracted for the final summary.

But the extracted sentences may or may not relate to each other.

Topic Identification
Based on word frequency
cannot detect all important information such as synonyms in the text

Redundancy
High keyword ranking can introduce redundancies in the summary.
The summary become concentrated around one specific topic.
Methods such as LSA can be used in order to reduce the amount of redundancy.

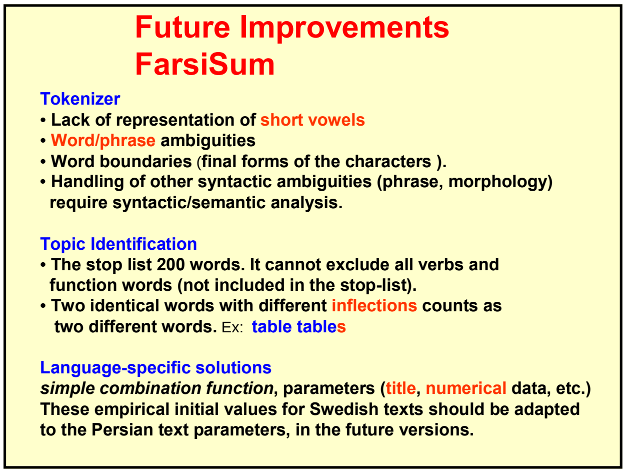


Notes on FarsiSum & SweSum

HTML Parser
frames, images, etc are not supported.
Missing *charset* causes problem.

Program Structure
- SweSum uses a plain structure
- English, French, German, Danish etc. in the same module make the program code unreadable and difficult to modify.
- Languages such as Persian using Unicode characters, should be in different modules

Programming language
Perl
very powerful & flexible script language for text management (tokenizing).
Syntax: regular expressions & data types

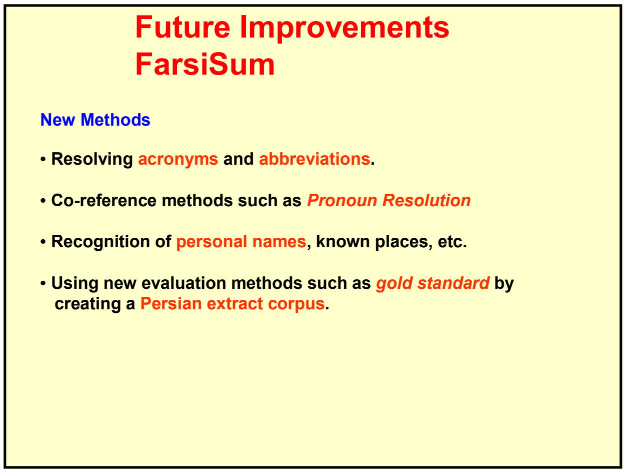


Future Improvements FarsiSum

Tokenizer
- Lack of representation of short vowels
- Word/phrase ambiguities
- Word boundaries (final forms of the characters).
- Handling of other syntactic ambiguities (phrase, morphology) require syntactic/semantic analysis.

Topic Identification
- The stop list 200 words. It cannot exclude all verbs and function words (not included in the stop-list).
- Two identical words with different inflections counts as two different words. Ex: table tables

Language-specific solutions
simple combination function, parameters (title, numerical data, etc.)
These empirical initial values for Swedish texts should be adapted to the Persian text parameters, in the future versions.



Future Improvements FarsiSum

New Methods

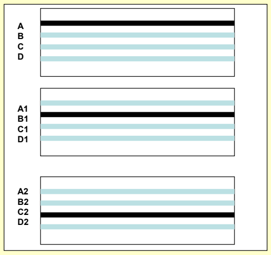
- Resolving acronyms and abbreviations.

- Co-reference methods such as *Pronoun Resolution*

- Recognition of personal names, known places, etc.

- Using new evaluation methods such as *gold standard* by creating a Persian extract corpus.

Future Improvements SweSum



Cohesion

Ordered list 1,2,3,4 → 2,4 in summary

Combine the linear structure used in SweSum with non-linear methods that operate on block level i.e.

collection of sentences rather than sentences

- to give higher score to lines adjacent to a line with high score
- Use OO structure in HTML, tags <P>
- Ordered list <OL> 1,2,3,.. Unordered list <UL> bullets

Future Improvements SweSum

HTML Parser

Increasing the *coherence* of the summarized text by

- Using HTML tags such as paragraph (<P>), Ordered List (<OL>), Unordered List (<UL>), etc
- The HTML tag <STRONG> should be handled as a <BOLD> get a higher score.
- Support for *frames* in the HTML code.
- Saving the *charset* parameter in the HTML *header*. It can be used in recovering of the encoding in case it is missing in the final summary.

Future Improvements SweSum

Program Structure

Solution I

The current program structure, but some units are improved:

1 Improvement of the parsing process.

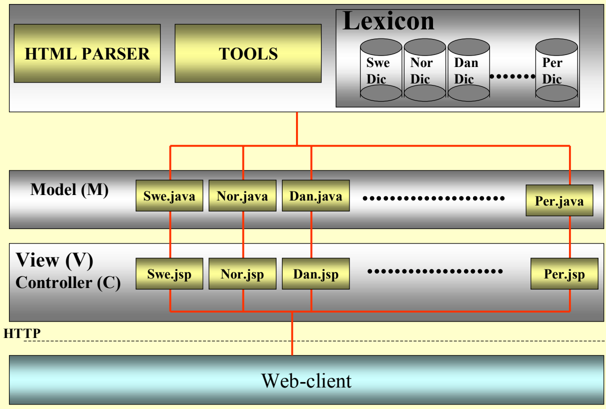
2 Each language has its own module or at least similar languages are in the same programming block. For example a partition between languages according to the used encoding (Latin1, Unicode, etc) is suitable.

Solution II

Using an external HTML parser.

Solution III

- Using an external HTML parser
- OO programming language such as Java, C++ or Object Oriented Perl
- Java is the best option since it has support for Unicode and provides a growing number of Internet tool resources such as Servlet, JavaBeans, JSP etc.



Conclusion

•As expected the field test showed that despite the ambiguity problems in Persian texts and use of a very simple stop-list, the final summary was improved both in the *coherence* and the preservation of *important information*.

•Use of an object oriented programming language which has support for Unicode, in the implementation of the future versions of SweSum is necessary.

•Tokenization process in languages using an Arabic writing system is different due to lack of representation of short vowels in the script and word/phrase ambiguities.

•Most of methods used in SweSum are applicable to Persian but in some cases language-specific solutions are required. For example the initial scoring values are empirical and language-dependent.

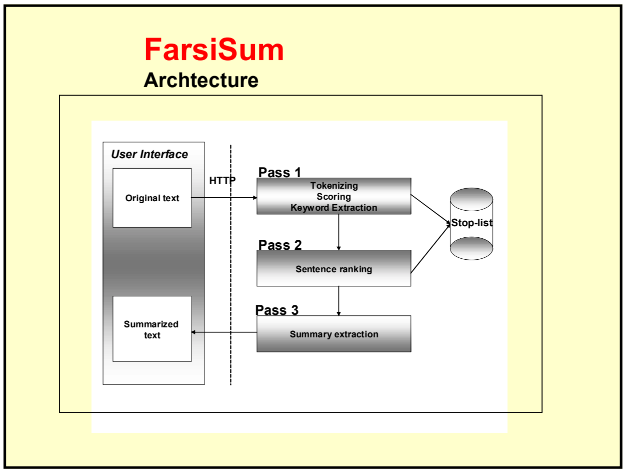
•To use co-reference methods such as *Pronoun Resolution*, *Synonym Resolution*, recognition of personal names, known places, etc in order to make the summarized text more coherent.

User Interface

Orginal text in UTF8 format

The user interface includes:

- The first page of FarsiSum on WWW presented in Persian.
  http://www.nada.kth.se/iplab/hlt/farsisum/index-farsi.html
- A Persian online editor for writing in Persian.
- The final summary including statistical information to the user, presented in Persian.



FarsiSum

Archtecture

User Interface

Original text

HTTP

Pass 1

Tokenizing
Scoring
Keyword Extraction

Stop-list

Pass 2

Sentence ranking

Pass 3

Summary extraction

Summarized text