FarsiSum - A Persian text summarizer

Martin Hassel
KTH NADA
Royal Institute of Technology
100 44 Stockholm, Sweden
xmartin@nada.kth.se

Nima Mazdak
Department of Linguistics
Stockholm University
106 91 Stockholm, Sweden
nima.mazdak@comhem.se

Abstract

FarsiSum is an attempt to create an automatic text summarization system for Persian. The system is implemented as a HTTP client/server application written in Perl. It uses modules implemented in an existing summarizer geared towards the Germanic languages, a Persian stop-list in Unicode format and a small set of heuristic rules.

1 Introduction

FarsiSum is an attempt to create an automatic text summarization system for Persian (Mazdak, 2004). The system is implemented as a HTTP client/server application written in Perl. It uses modules implemented in SweSum (Dalianis 2000), a Persian stop-list in Unicode format and a small set of heuristic rules. The stop-list is a file including the most common verbs, pronouns, adverbs, conjunctions, prepositions and articles in Persian. The words not included in the stop-list are supposed to be nouns or adjectives. The idea is that nouns and adjectives are meaning-carrying words and should be regarded as keywords.

The current implementation of FarsiSum is still a prototype. It uses a very simple stop-list in order to filter and identify the important keywords in the text. Persian acronyms and abbreviations are not detected by the current tokenizer.

In addition, Persian syntax is quite ambiguous in its written form (Megerdoomian and Rémi 2000), which raises certain difficulties in automatic parsing of written text and automatic text summarization for Persian.

For example, selection of important keywords in the *topic identification* process will be affected by the following word boundary ambiguities:

- Compound words may appear as two different words.
- Bound morphemes may appear as free morphemes or vice versa.

These ambiguities are not resolved in the current implementation.

2 SweSum

SweSum[1] (Dalianis 2000) is a web-based automatic text summarizer developed at the Royal Institute of Technology (KTH) in Sweden. It uses text extraction based on statistical and linguistic as well as heuristic methods to obtain text summarization and its main domain is Swedish HTML-tagged newspaper text[2].

2.1 SweSum's architecture

SweSum is a client/server application. The summarizer is located on the web server. It takes a Swedish text as input and performs summarization in three phases to create the final output (the summarized text).

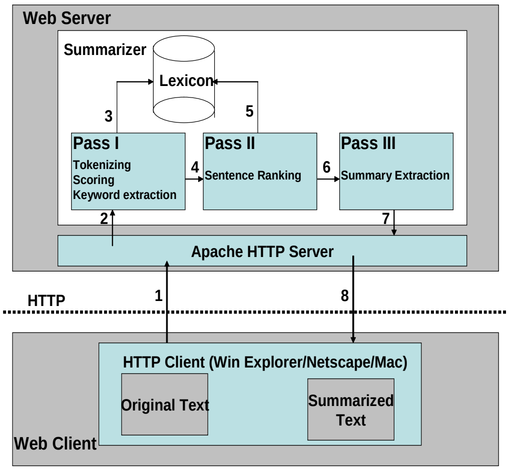

Figure 1: SweSum architecture

Pass 1: The sentence and word boundaries are identified by searching for periods, exclamation and question marks etc (with the exception of when periods occur in known abbreviations). The sentences are then scored by using statistical, linguistic and heuristic methods. The scoring depends on, for example, the position of the sentence in the text, numerical values in and

[1] An online demo is available at http://swesum.nada.kth.se/index.html

[2] SweSum is also available for English, Danish, Norwegian, Spanish, French, German, and now with the implementation described in this paper, Farsi.

various formatting of the sentence such as bold, headings, etc.

Pass 2: In the second pass, the score of each word in the sentence is calculated and added to the sentence score. Sentences containing common content words get higher scores.

Pass 3: In the third pass, the final summary file (HTML format) is created. This file includes:

- The highest ranking sentences up to a pre-set threshold.
- Optionally, statistical information about the summary, i.e. the number of words, number of lines, the most frequent keywords, actual compression rate etc.

For most languages SweSum uses a static lexicon containing many high frequent open class words. The lexicon is a data structure for storing key/value pairs where the key is the inflected word and the value is the stem/root of the word. For example *boy* and *boys* have different inflections but the same root (lemma).

3 FarsiSum

FarsiSum is a web-based text summarizer for Persian based upon SweSum. It summarizes Persian newspaper text/HTML in Unicode format. FarsiSum uses the same structure used by SweSum (see Figure 2), with exception of the lexicons, but some modifications have been made in SweSum in order to support Persian texts in Unicode format.

3.1 User Interface

The user interface includes:

- The first page of FarsiSum on WWW presented in Persian[3].
- A Persian online editor for writing in Persian.

The final summary including statistical information to the user, presented in Persian.

3.2 Stop List

The current implementation uses a simple stop list rather than a full-fledged Persian lexicon. The stop-list is a HTML file (UTF-8 encoding) containing about 200 high-frequency Persian words including the most common verbs, pronouns, adverbs, conjunctions, prepositions and articles. The stop-list has been successively built during the implementation phase by iteratively running FarsiSum in order to find the most common words in Persian.

The assumption is that words not included in the stop-list are nouns or adjectives (content words) and should be counted as such in the word frequency list.

3.3 Tokenizer

The tokenizer is modified in order to recognize Persian comma, semi colon and question mark.

- Sentence boundaries are found by searching for periods, exclamation and question marks as well as <BR> (the HTML new line) and the Persian question mark (؟).
- The tokenizer finds the word boundaries by searching for characters such as “.”, “,”, “!”, “?”, “<”, “>”, “:”, spaces, tabs and new lines. Persian semi colon, comma and question mark can also be recognized.
- All words in the document are converted from ASCII to UTF-8. These words are then compared with the words in the stop-list. Words not included in the stop list are regarded as content words and will be counted as keywords.

The word order in Persian is SOV[4], i.e. the last word in a sentence is a verb. This knowledge is used to prevent verbs from being stored in the *Word frequency table*.

3.4 Architecture

FarsiSum is implemented as a HTTP client/server application as shown in Figure 2. The summarization program is located on the server side and the client is a browser such as Internet Explorer or Netscape Navigator.

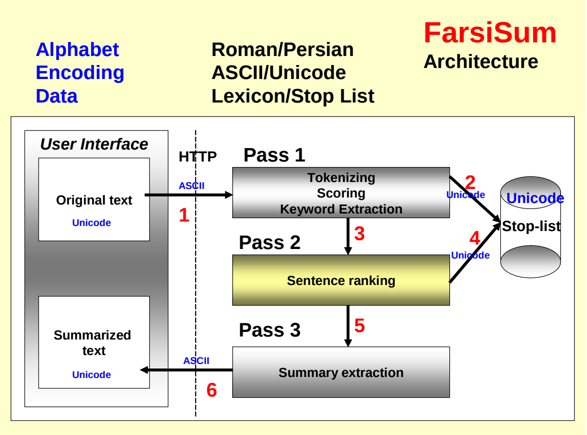

Figure 1: FarsiSum architecture

[3] http://www.nada.kth.se/iplab/hlt/farsisum/index-farsi.html

[4] **SOV** stands for **S**ubject, **O**bject and **V**erb.

The summarization process starts when the user (client) clicks on a hyperlink (*summarize*) on the FarsiSum Web site:

- The browser (Web client) sends a summarization request (marked 1 in Figure 2) to the Web server where FarsiSum is located. The document/ (URL of the document) to be summarized is attached to the request. (The original text is in Unicode format).
- The document is summarized in three phases including tokenizing, scoring and keyword extraction. Words in the document are converted from ASCII to UTF-8. These words are then compared with the words in the stop-list (2-5).
- The summary is returned back to the HTTP server that returns the summarized document to the client (6).

The browser then renders the summarized text to the screen.

4 Conclusions

The system would most certainly benefit from deeper language specific analysis, but with no access to Persian resources, in this system fairly language independent methods have proven to come a long way.

References

Dalianis, H. 2000. *SweSum - A Text Summarizer for Swedish*, Technical report, TRITA-NA-P0015, IPLab-174, NADA, KTH, October 2000.

Mazdak, N. 2004. *FarsiSum - a Persian text summarizer*, Master thesis, Department of Linguistics, Stockholm University, (PDF)

Megerdoomian, Karine and Rémi, Zajac 2000. *Processing Persian Text: Tokenization in the Shiraz Project*. NMSU, CRL, Memoranda in Computer and Cognitive Science (MCCS-00-322).