Towards Building a WordNet for Persian Adjectives

Ali Famian
Linguistics Department
Tarbiyat Modares University
Tehran, Iran,
FamianAli@Yahoo.Com

Daruosh Aghajaney
Computer Department
Jahad Higher Education Institute
Tehran, Iran
Daruosha@Daruosha.Com

Abstract

This article attempts to report on a project for building a WordNet for Persian adjectives. Three monolingual Persian dictionaries, as well as a Farsi linguistic corpus are employed here to extract required entries. This WordNet provides the semantic classes of adjectives, their synonyms, antonyms and frequency. The database management system employed here is MS SQL Server 2000. The system is implemented using Microsoft's .NET Framework in Visual C# language and is developed in both desktop and web-based platforms. It allows the end-user to export the results of a specific query, both in Persian and Latinized alphabets, into a CSV or XML file for further reference.

1 Introduction

The growing importance of Natural Language Processing tools made us start building a WordNet for Persian adjectives. Our work is structured mainly on the principles of the original Princeton WordNet. We will also work along the line of PersiaNet, a project for developing a WordNet for Persian nouns and then verbs. In this paper, first, we introduce our distinguished semantic classes, and also our selected lexical resources. Next an overview on the user interface of the system and its features and functionalities is given. The database architecture is also presented with some details on main issues and the relations between tables. The final section concerns the platform and development environment of our WordNet.

2 Adjectives and Lexical Resources

In this project, we identify different classes of Persian adjectives. Following GermaNet approach, 15 main classes are distinguished first, and then they are sub-divided further into more specific ones. Persian Adjective WordNet aims to cover around 5000 adjective entries. To extract the data, a combination of manual and automatic methods is used. To do so, besides WordNet 2.0, we employ three monolingual Persian dictionaries, as well as an electronic database. Anvari (2000) and Sadri Afshar (1988) are our first and main resources to select the adjectives and their semantic classes. Then we use Khodaparasti (1997), a comprehensive dictionary of Persian synonyms and antonyms to relate the lexical items. As an electronic Persian corpus, Assi (1997) provides a means of handling various types of texts to determine the frequency of adjectives.

2.1 Graphic User Interface (GUI)

The GUI supports the following functionalities:

2.1.1 Querying

- Searches entries with specific information
- Extracts statistic information from the data (the number of relations, the semantic class of adjective, etc.)

2.1.2 Viewing the data

- Individual senses
- Browsing through entries and links

2.1.3 Exporting

- Exports all or a specific entry into a Comma Separated Value (CSV) and eXtended Meta Language (XML) file.

This interface supports two character sets i.e. Persian and Latinized alphabets. By entering some part of a given entry, the system offers the potential similar words.

3 Database Architecture

The database is implemented using Microsoft^{TM} SQL Server 2000, including the following tables:

1. Entries Table: Each entry is entered in one record.
2. Descriptions Table: It contains descriptions on records of entries table. Each entry may have more than one descriptions.
3. Synonyms Table: The similar descriptions of entries point to each other.
4. Antonyms Table: The same as the Synonyms Table.
5. Labels Table: The semantic classes of adjectives are listed here with relation to Descriptions Table. Two levels are considered for labeling the semantic classes of adjectives : the first level classifies the entiries into 15 catagories, and then they are sub-divided into over 70 more specific groups.

In Entries, Descriptions, and Labels tables, the data are entered in both Persian and Latinized alphabet system.

* The figure at the end of the article represents the database arcitecture diagram.

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 307–308.

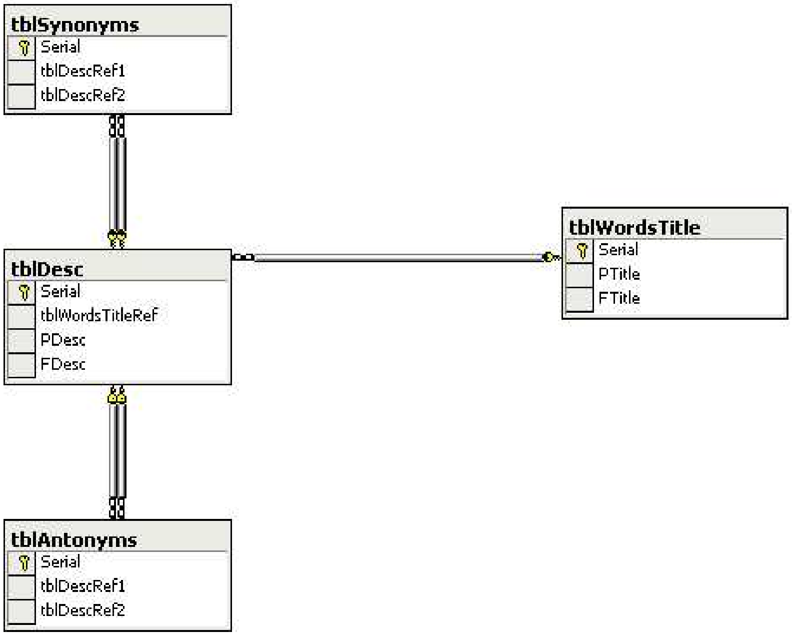

Figure 1: Database Architecture Diagram

4 Platform and Development Environment

The Persian Adjectives WordNet is developed using Microsoft^{TM} Visual Studio 2003 with C# language. The application is developed in both desktop and web-based platforms. The database connection of the applications can be configured by the end-user. The administration panel is built in the desktop format with high level of security and ease. Using the web-based technology, the end-user is free to choose the environment and operating systems, e.g. Linux, Sun^{TM} Solaris, Fedora, Microsoft^{TM} Windows, etc.

Conclusion

In this article, we described the on-going work on creating a WordNet for Persian adjectives. In the process of project development, the semantic classification of Persian adjectives revealed stunning facts on Persian semantics.This WordNet covers around 5000 entries, and the response time for each query with the average level of relations (synonyms, antonyms, as well as descriptions) is very close to real time. Due to our logical design of database, we hope to link our wordnet to the PersiaNet system, a WordNet for Persian nouns and verbs, developed at Princeton University.

Acknowledgements

We have received helpful comments from Dr. Mostafa Assi, especially on the electronic Persian corpus developed at The Center for Humanities and Cultural Studies, Tehran, Iran. And special thanks are due to Dr. M. M. Vahedi Langrudi, for his everlasting help and support.

References

Anvari, H. (2004) *Sokhan Dictionary* (2 Vol.), Tehran: Sokhan Publishers. 2704 p.

Assi, S. M. (1997) *Farsi Linguistic Database (FLDB*), International Journal of Lexicography, Vol. 10, No. 3, Euralex Newsletter.

Cruse, D. A. (1986) *Lexical Semantics* Cambridge, Cambridge University Press. 310 p.

Dixon, R. M. W. (1982) Where Have All the Adjectives Gone? In: Robert M.W. Dixon, Where Have All the Adjectives Gone? and Other Essays in Semantics and Syntax. Berlin-Amsterdam-New-York: Mouton. pp. 1–62.

Fellbaum, C. (1995) "Co-occurrence and antonymy." In: International Journal of Lexicography 8 (4), pp. 281–303.

Fellbaum, C. (Ed.) 1998) "*WordNet: An Electronic Lexical Database*". The MIT Press, Cambridge. 423 p.

Hamp, B. and Feldweg, H. (1997) "*GermaNet - a Lexical-Semantic Net for German*". In: Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications". Madrid.

Khodaparasti, F. (1997) *A Comprehensive Dictionary of Persian Synonyms and Antonyms*, Shiraz: Daneshnameye Fars. 488 p.

Miller, G, R. Beckwith, C. Fellbaum, D. Gross and K. Miller (1990) *Five papers on WordNet*, CSL Report 43. Cognitive Science Laboratory. Princeton University, 89 p.

Patton, Robert and Ogle J. (2001) *Designing SQL Server 2000 Databases for .NET Enterprise Servers.* Syngress Publishing, Inc., 753 p.

Raskin, v., and Nirenburg S. (1995) Lexical Semantics of Adjectives, A Microtheory of Adjectival Meaning, MCCS Report pp. 95–288.

Sadri Afshar, et al. (1998) *A Dictionary of Persian Language, Tehran*: Kaleme Publisher, 846 p.