Urdu Localization Project: Lexicon, MT and TTS (ULP)

Sarmad HUSSAIN
Center for Research in Urdu Language Processing,
National University of Computer and Emerging Sciences
B Block, Faisal Town
Lahore, Pakistan
sarmad.hussain@nu.edu.pk

Abstract

Pakistan has a population of 140 million speaking more than 56 different languages. Urdu is the lingua franca of these people, as many speak Urdu as a second language, also the national language of Pakistan. Being a developing population, Pakistani people need access to information. Most of the information over the ICT infrastructure is only available in English and only 5-10% of these people are familiar with English. Therefore, Government of Pakistan has embarked on a project which will generate software to automatically translate the information available in English to Urdu. The project will also be able to convert Urdu text to speech to extend this information to the illiterate population as well. This paper overviews the overall architecture of the project and provides briefs on the three components of this project, namely Urdu Lexicon, English to Urdu Machine Translation System and Urdu Text to Speech System.

1 Introduction

In today's information age it is critical to provide access to information to people for their development. One precursor to this access is availability of information in the native languages. Due to limitations in technology, it has not been possible to generate information in many languages of the world. However, with recent advances in internationalization and localization technology, many languages are not enabled. However, as this is recent development, the published content in these languages is still limited, and far lags behind the content available for English, Spanish and some other languages spoken in developed countries. Realizing this gap in content and the need to provide access to information to its citizens, Government of Pakistan has recently launched Urdu Localization Project[1]. This project will enable translation and access of English content to literate and illiterate Urdu speakers.

Urdu Localization Project aims to provide access to existing English language content to Urdu language speakers. The project has three components: Urdu Computational Lexicon, English-to-Urdu Machine Translation System, Urdu Text-to-Speech system. This paper briefly describes the architecture and work achieved to-date for different systems within ULP.

2 ULP Architecture

As indicated, ULP comprises of three largely independent systems: Lexicon, MT and TTS, though these components may also be integrated to develop a written and oral interface to information. The project has three architectural layers. At the base are the core data and engines for Lexicon, MT and TTS. The middle layer provides public programming interfaces to these engines (APIs) so that they may be integrated with end-user applications at the top layer or used by third-party applications. Both the engine and API layer components are being developed in standard C/C++ to enable them to compile on all platforms (e.g. Microsoft, Linux, Unix). The user-end/top layer has to be technology centric and is currently being enabled in Microsoft platform. The lexicon will be given a web interface for user access. In addition, plug-ins for internet and email clients will be developed for MT and TTS to enable end-users to translate and re-display English websites in Urdu and also enable them to convert the translated Urdu text to speech. This is shown in Figure 1 below. In the figure the layers and systems are demarcated (horizontally and vertically respectively). The figure also shows that MT and TTS may be using the Lexicon through the APIs for getting appropriate data.

[1] Urdu Localization Project is a three-year initiative being undertaken by Center for Research in Urdu Language Processing (www.crulp.org) and is funded through a grant by E-Government Directorate of Ministry of IT&Telecom., Government of Pakistan.

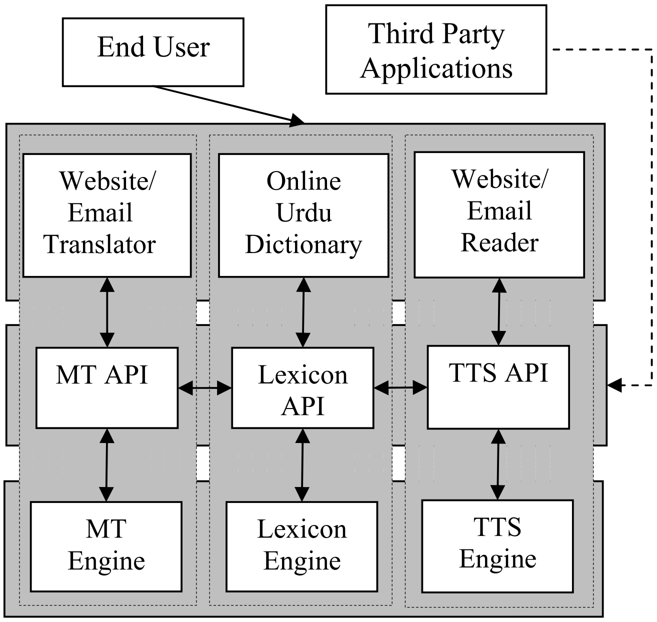

Figure 1: Architecture Diagram for ULP

These three systems are discussed briefly below.

2.1 Urdu Lexicon

Urdu Computational Lexicon being designed would be holding more than 25 dimensions of a single headword. The first task to date has been to determine this hierarchical storage structure. The structure required for end-user has been finalized. However, requirements for computational applications, e.g. MT, are still being finalized. This was perhaps one of the most challenging tasks as there are currently no standards which exist, although some guidelines are available. In addition, Urdu also had some additional requirements (e.g. multiple plural forms, depending on whether the word is derived from Arabic or Sanskrit). Entries of more than thirty thousand headwords and complete entry of about a thousand headwords along with specification of at least 15 entries has already been done. Currently more content is being generated. In addition, work is under progress to define the physical structure of the lexicon (e.g. storage and retrieval models). The prototype showing this application is also available in Microsoft platform.

2.2 English-Urdu Machine Translation

Work is under progress to develop English to Urdu MT engine. The translation is based on LFG formalism and is developing grammars, lexica and the parsing/mapping/generation engine for LFG. Mapping and Generation prototypes have already been developed and are integrated with a freely available LFG parser for internal testing. In addition sample grammars for English, Urdu and English-Urdu mapping have also been written. The prototype covers about 10 percent of grammatical rules and already translates within the limited vocabulary of the engine. The work is being extended to write the parser and rewrite mapper and generator and to develop English, Urdu and English Urdu grammars and lexica.

2.3 Urdu Text to Speech System

The Urdu TTS is divided into two main part, the Urdu Natural Language Processor and Urdu Speech Synthesizer. The work on NLP is completed (except the intonational module, on which preliminary work has been completed). The NLP processor inputs Urdu Unicode text and output narrow phonetic transcription with syllable and stress markers. The NLP processor is integrated with Festival speech synthesis system (though by-passes its NLP module). A vocabulary of about 500 words is already defined at the diphones have been created. Prototype application is already developed which synthesized these single words. Work is currently in progress to define Urdu intonational and durational model. In addition, work is also under progress to extend the vocabulary and functionality to synthesize complete sentences. The functional prototype works on both Linux an Microsoft platforms.

3 Conclusion

Most of the work being done in the project is novel. Urdu language is not very well defined for use with computers. Script, speech and language aspects of Urdu are being studied, documented and implemented in this project. The project is also testing the work which has been matured on western languages but only being recently exposed to other languages, e.g. the lexical recommendations by ISLE, LFG framework, use of LFG for MT, speech modeling of Urdu (both spectral and temporal) and more. Non-functional issues including performance are also being negotiated. Pre-compiled lexica, user-centric pre-stored performance-enhancing profiles and frequency lists, etc. are part of the architectural tasks being addressed. Though only initial work has been done, this work in itself is substantial, and has raised many questions which will be answered as the project progresses.