Resources for Urdu Language Processing

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

Urdu is spoken by more than 100 million speakers. This paper summarizes the corpus and lexical resources being developed for Urdu by the CRULP, in Pakistan.

1 Introduction

Urdu is the national language of Pakistan and one of the state languages of India and has more than 60 million first language speakers and more than 100 million total speakers in more than 20 countries (Gordon 2005) . Urdu is written in Nastalique writing style based on Perso-Arabic script. This paper focuses on the Urdu resources being developed, which can be used for research in computational linguistics.

2 Urdu Text Encoding

Urdu computing started early, in 1980s, creating multiple encodings, as a standard encoding scheme was missing at that time. With the advent of Unicode in early 1990s, some online publications have switched to Unicode, but much of the publication still continues to follow the ad hoc encodings (Hussain et al. 2006). Two main on-line sources of Urdu text in Unicode are Jang News (www.Jang.net/Urdu) and BBC Urdu service (www.BBC.co.uk/Urdu) and are thus good sources of corpus. Encoding conversion may be required if data is acquired from other sources.

3 Corpora

EMILLE Project, initiated by Lancaster University is one of the first initiatives to make Urdu corpus available for research and development of language processing (McEnery et al. 2000). The project has released 200,000 words of English text translated into Bengali, Gujarati, Hindi, Punjabi and Urdu, creating a parallel corpus across these languages. In addition, the corpus also has 512,000 words of Spoken Urdu, from BBC Radio. Moreover, the corpus also contains 1,640,000 words of Urdu text. These Urdu corpus resources are also annotated with a large morpho-syntactic tag-set (Hardie 2003).

Center for Research in Urdu Language Processing (CRULP) at National University of Computer and Emerging Sciences in Pakistan has also been developing corpora and associated tools for Urdu. A recent project collected a raw corpus of 19 million words of Urdu text mostly from Jang News, reduced to 18 million words after cleaning. The corpus collection has been based on LC-STAR II guidelines[1]. The domain-wise figures are given in Table 1. Further details of the corpus and associated information are discussed by Ijaz et al. (2007).

Table 1: Distribution of Urdu Corpus

Domains	Cleaned Corpus	
	Total Words	Distinct Words
C1. Sports/Games	1529066	15354
C2. News	8425990	36009
C3. Finance	1123787	13349
C4. Culture/Entertainment	3667688	34221
C5. Consumer Information	1929732	24722
C6. Personal communications	1632353	23409
Total	18308616	50365

Agreement between CRULP and Jang News allows internal use. However, due to distribution restrictions in this agreement, the corpus has not been made publicly available. The distribution rights are still being negotiated with Jang News.

The tag set developed by Hardie (2003) is based on morpho-syntactic analysis. A (much reduced) syntactic tag set has also been developed by

[1] See www.lc-star.org/docs/LC-STAR_D1.1_v1.3.doc

CRULP (on the lines of PENN Treebank tagset), available at its website www.CRULP.org. A corpus of 100,000 words manually tagged on this tag set has also been developed based on text from Jang online news service. This *CRULP POS Tagged Jang News Corpus* is available through the center.

Recently another corpus of about 40,000 words annotated with Named Entity tags was also made available for Workshop on NER for South and South East Asian Languages organized at IJCNLP 2008. The annotated corpus was donated by CRULP and IIIT Hyderabad and is available at http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=5. Tag set contains 12 tags. Details of these tags are discussed at the link http://ltrc.iiit.ac.in/ner-ssea-08/index.cgi?topic=3. The CRULP portion of the data is also available at CRULP website, and is a subset of the *CRULP POS Tagged Jang News Corpus.*

In earlier work at CRULP, a 230 spelling errors corpus has also been developed based on typographical errors in Newspapers and student term papers. See Naseem et al. (2007) for details.

A corpus of Urdu Names has also been developed by CRULP, based on the collective telephone directories of Pakistan Telecommunications Corporation Limited (PTCL) from across all major cities of Pakistan. A name list has also been extracted from the corpus for all person names, addresses and cities of Pakistan.

4 Lexica

Lexica are as critical for development of language computing as corpora. One of the most comprehensive lexica available for Urdu was recently released by CRULP (available through CRULP website). The online version, called Online Urdu Dictionary (OUD) contains 120,000 entries, with 80,000 words annotated with significant information. The data of OUD is XML tagged, as per the annotation schema discussed by Rahman (2005; pp. 15), which contains about 20 etymological, phonetic, morphological, syntactic, semantic and other parameters of information about a word. The dictionary also gives translation of 12000 words in English and work is under way to enable runtime user-defined queries on the available XML tags. The contents of this lexicon are based on the 21 volume Urdu Lughat developed by Urdu Dictionary Board of Government of Pakistan. See www.crulp.org/oud for details.

CRULP has also developed a corpus based lexicon of 50,000 words with frequency data and annotation specifications defined by LC-STAR II project (at http://www.lc-star.org/docs/LC-STAR_D1.1_v1.3.doc). Details of the lexicon annotation scheme are given by Ijaz et al. (2007).

There are also additional tools available through CRULP, and documented at its website, including normalization, collations, spell checking, POS tagging and word segmentation applications.

5 Conclusions

This paper lists some core linguistic resources of Urdu, available through CRULP and other sources. However, the paper identifies licensing constraints, a challenge for open distribution, which needs to be addressed.

References

Gordon, Raymond G., Jr. (ed.). (2005). *Ethnologue: Languages of the World*, Fifteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/.

Hardie, A. (2003). Developing a tag-set for automated part-of-speech tagging in Urdu. In *Archer, D, Rayson, P, Wilson, A, and McEnery, T (eds.) Proceedings of the Corpus Linguistics 2003 conference. UCREL Technical Papers Volume 16.* Department of Linguistics, Lancaster University, UK.

Ijaz, M. and Hussain, S. (2007). Corpus Based Urdu Lexicon Development. In the *Proceedings of Conference on Language Technology '07*, University of Peshawar, Peshawar, Pakistan.

Naseem, T. and Hussain, S. (2007). Spelling Error Trends in Urdu. In the *Proceedings of Conference on Language Technology '07*, University of Peshawar, Peshawar, Pakistan.

McEnery, A., Baker, J., Gaizauskas, R. & Cunningham, H. (2000). EMILLE: towards a corpus of South Asian languages, *British Computing Society Machine Translation Specialist Group*, London, UK.

Rahman, S. (2005). Lexical Content and Design Case Study. Presented at *From Localization to Language Processing, Second Regional Training of PAN Localization Project.* Online presentation version: http://panl10n.net/Presentations/Cambodia/Shafiq/LexicalContent&Design.pdf.