Abs-34

Mohammad Abid Khan and Fatima Tuz Zuhra (University of Peshawar, Pakistan)

Role of Corpus in Anaphora Resolution

Human beings use pronouns for avoiding repetitions but computers need to process text in simplified form for a number of potential applications such as machine translation. A key component of text simplification is anaphora resolution e.g. replacing pronouns by their antecedents. In this research paper, a corpus-based approach for resolving anaphora is worked out. Results show that this approach is much more economical compared to traditional strategies. Pashto corpus is divided into two parts: training part and testing part. Anaphora resolution rules are learnt automatically from the training part and subsequently tested on the testing part of the corpus. Both parts are carefully selected as representative samples from the Pashto corpus of 1.225 million words (Khan and Zuhra, 2009). Discourse boundaries are identified in the training part of the corpus having 14000 words. There are 68 discourse units in the training sample. Anaphora is resolved manually in the training part and the anaphoric and non-anaphoric versions of each discourse unit are stored side by side with each other. This text is then part-of-speech (POS) tagged using the POS tagger for Pashto (Khan, 2010a) including numbering of nouns and noun phrases. Tagged sequences are extracted from the text in such a way that the anaphoric and the non-anaphoric sequences of each discourse unit are saved parallel to each other in a Microsoft Access database table. Later, tagged sequences of sentences are extracted from both versions of the tagged sequences of discourse units using the methodology of Khan (2010b). The tagged sequence of each anaphoric sentence is compared with the tagged sequence of its non-anaphoric counterpart. The differences are recorded. These differences form the basis for extracting the anaphora resolution rules. A total 492 sentence pairs of the training part are compared and rules are formulated automatically. These rules are then manually normalized using observation of the tagged sequences of discourse units and the corresponding text. A total 142 anaphora resolution rules are bagged in this process. These rules are tested in a semi-automatic way using a testing sample of 50 discourse units containing 371 pronouns, out of which 59 pronouns occurred in a non-anaphoric way i.e. used for pointing purposes. Of the remaining 312 pronouns, 230 pronouns are correctly resolved using the rules learnt by the system. This is the first step in this direction. Now the size of the training part of the corpus can be increased by adding to it the automatically resolved text and new text is added to the testing corpus. Repetition of this iterative process will very quickly lead to the perfection of the technique.

References

Baker, P. et al. 2008. “Discourse and Society”.[Online]. Available from the URL: http://das.sagepub.com/cgi/content/abstract/19/3/273

Ali, R., Khan, M.A. and Rabbi, I., “Strong Personal Anaphora Resolution in Pashto Discourse”, In proc. IEEE ICET 3rdInternational Conference on Emerging Technologies, Islamabad, Pakistan. 2007, pp 148-154.

Ali, R., Khan, M.A. and Rabbi, I., “Reflexive Anaphora Resolution in Pashto Discourse”, In proc. Conference on Language and Technology, 2008.

Gasperin, C. et al. 2009. Learning When to Simplify Sentences for Natural Text Simplification. In the proc. ENIA 2009 (VII EncontroNacional de Inteligencia Artificial), Bento Goncalvez, RS, Brazil.

Jurafsky, D. and Martin, J. H. 2002.Speech and Language Processing. Pearson Education Series in Artificial Intelligence, Colorado.

Khan, M. A. 2010 a. “A Part of Speech Tagger for Pashto”, Working paper, Department of Computer

Science, University of Peshawar, 2010.

Khan, M. A. 2010 b. “Extraction of Grammar Rules from the Pashto Corpus”, Working paper, Department of Computer Science, University of Peshawar, 2010.

Khan, M. A. and Zuhra. 2007. “A General-Purpose Monitor Corpus of Written Pashto”. In proc. Corpus Linguistics 2007.Birmingham, UK.

Khan, M. A. and Zuhra. 2009. “A Corpus-Based Study of Pashto”. In proc. Corpus Linguistics 2009. Liverpool, UK.

McEnery, T. et al. 2006.Corpus-Based Language Studies: An Advanced Resource Book. Routledge Applied Linguistics.

Ooi, V. 2001. Investigating and Teaching Genres Using the World Wide Web”. In M. Ghadessy, A. Henry and R.L. Roseberry (eds) Small Corpus Studies and ELT. Amsterdam: Benjamins.