Building Computational Resources: The URDU.KON-TB Treebank and the Urdu Parser

By: Qaiser Abbas **Supervisor:** Prof. Dr. Miriam Butt

Abstract

This work presents the development of the URDU.KON-TB treebank, its annotation evaluation & guidelines and the construction of the Urdu parser for a South Asian language Urdu. Urdu is comparatively an under-resourced language and the development of a reliable treebank and a parser will have significant impact on the state-of-the-art for automatic Urdu language processing.

The work includes the construction of the raw corpus containing 1400 sentences collected from Urdu Wikipedia and the Jang newspaper. The corpus contains text of local & international news, social stories, sports, culture, finance, religion, traveling, etc. The hierarchal annotation scheme adopted has a combination of phrase structure and hyper dependency structure. A semi-semantic part of speech tag set, a semi-semantic syntactic tag set and a functional tag set are proposed, which are further revised during the annotation of the raw corpus. The annotation of the sentences was performed manually. Due to the addition of morphology, part of speech, syntactical, semantical, clausal, grammatical and miscellaneous features, the annotation scheme is linguistically rich. The annotation resulted in a treebank for Urdu, called the URDU.KON-TB.

For an evaluation of the annotation scheme, Krippendorff's α co-efficient is selected. This is a statistical measure to evaluate inter-annotator agreement. Randomly selected 100 sentences from the URDU.KON-TB treebank were given to five trained annotators for annotation. The annotated sentences were then evaluated using the Krippendorff's α co-efficient. The percentage values of inter-annotator agreement obtained for part of speech, syntactical and functional annotations are 96%, 82% and 81%, respectively. All of the three values lie in the range of perfect agreement. The annotation guidelines devised in the development of the URDU.KON-TB treebank were revised during and after this annotation evaluation.

For the development of an Urdu parser, 1400 annotated sentences in the URDU.KON-TB treebank are divided into 80% training data and 20% test data. A context free grammar is extracted from this training data, which is then given to the Urdu parser after its development. The test data is divided into 10% held out data and 10% test data. The test data then contains 140 sentences with an average length of 13.73 words per sentence. The held out data is used during the development of the Urdu parser. Urdu parser is an extended version of dynamic programming algorithm known as the Earley parsing algorithm. The

extensions made are discussed in doctoral thesis along with the issues faced during the development. All items which can occur in a normal text are considered, e.g., punctuation, null elements, diacritics, headings, regard titles, Hadees (the statements of prophets), anaphora with in a sentence, and others. The PARSEVAL measures are used to evaluate the results of the Urdu parser. By applying a sufficiently rich grammar along with the extended parsing model, the parser gives 87% of f-score and outperforms the multi-path-shift-reduce parser for Urdu, a two stage Hindi dependency parser and a simple Hindi dependency parser with 4.8%, 12.48% and 16.4% increase in recall, respectively.

The URDU.KON-TB treebank and the Urdu parser is a contribution to the overall computational resources of Urdu. By-products of this work are a semi-semantic part of speech tag set, a semi-semantic syntactic tag set, a functional tag set, annotation guidelines, a grammar with sufficient encoded information for parsing of morphologically rich language Urdu and a part of speech tagged corpus, which can be used for the training of part of speech taggers. These resources will be enhanced further and can be used for natural language processing such as probabilistic parsing, training of POS taggers, disambiguation of spoken sentences, grammar development, language identification, sources for linguistic inquiry and psychological modeling, or pattern matching.