

PREFACE

WORKSHOP THEME

Recently, there has been a surge of interest in the study of the languages of the Middle East, especially Arabic, Persian (Farsi), Pashto, Kurdish and Urdu. This sudden and urgent interest is manifested by the availability of funding for rapid development of practical systems for processing large volumes of data in these languages. Computational applications for proper name identification, entity recognition, categorization, information retrieval, summarization, machine translation and other implementations are currently in high demand. This comes at a time when advances in formal and computational linguistics over the last fifty years are being consolidated, while work on machine learning and statistical methods has been showing great promise.

There exists a considerable body of work in computational linguistics specifically targeted to these middle eastern languages. Much of the research and development has been the result of initiatives by individual research establishments or industry firms. Furthermore, the usage of the Arabic script gives rise to certain issues that are common to all these languages despite their being of distinct language families. Hence, these languages share properties such as the absence of capitalization, right to left direction, lack of clear word boundaries, complex word structure, a high degree of ambiguity due to non-representation of short vowels in the writing system, and related encoding issues.