OVERVIEW OF THE CLEF 2008 MULTILINGUAL QUESTION ANSWERING TRACK

Pamela Forner[1], Anselmo Peñas[2], Eneko Agirre[3], Iñaki Alegria[4], Corina Forăscu[5], Nicolas Moreau[6], Petya Osenova[7], Prokopis Prokopidis[8], Paulo Rocha[9], Bogdan Sacaleanu[10], Richard Sutcliffe[11], and Erik Tjong Kim Sang [12]

[1] CELCT, Trento, Italy (forner@celct.it)
[2] Departamento de Lenguajes y Sistemas Informáticos, UNED, Madrid, Spain (anselmo@lsi.uned.es)
[3]Computer Science Department, University of Basque Country, Spain (e.agirre@ehu.es)
[4]University of Basque Country, Spain (i.alegria@ehu.es)
[5] UAIC and RACAI, Romania (corinfor@info.uaic.ro)
[6] ELDA/ELRA, Paris, France (moreau@elda.org)
[7] BTB, Bulgaria, (petya@bultreebank.org)
[8] ILSP Greece, Athena Research Center (prokopis@ilsp.gr)
[9] Linguateca, DEI UC, Portugal, (Paulo.Rocha@di.uminho.pt)
[10] DFKI, Germany, (bogdan@dfki.de)
[11] DLTG, University of Limerick, Ireland (richard.sutcliffe@ul.ie)
[12] University of Groningen (e.f.tjong.kim.sang@rug.nl)

Abstract The QA campaign at CLEF [1], was manly the same as that proposed last year. The results and the analyses reported by last year's participants suggested that the changes introduced in the previous campaign had led to a drop in systems' performance. So for this year's competition it has been decided to practically replicate last year's exercise.
Following last year's experience some QA pairs were grouped in clusters. Every cluster was characterized by a topic (not given to participants). The questions from a cluster contained co-references between one of them and the others. Moreover, as last year, the systems were given the possibility to search for answers in Wikipedia[1] as document corpus beside the usual newswire collection.
In addition to the main task, three additional exercises were offered, namely the Answer Validation Exercise (AVE), the Question Answering on Speech Transcriptions (QAST), which continued last year's successful pilot, and Word Sense Disambiguation for Question Answering (QA-WSD).
As general remark, it must be said that the task still proved to be very challenging for participating systems. In comparison with last year's results the Best Overall Accuracy dropped significantly from 41,75% to 19% in the multi-lingual subtasks,

[1] http://wikipedia.org