

CNGL Undergraduate Students as Researchers Programme

PROJECT DESCRIPTION

Institution/Track:	Dublin City University	
Project Title:	Dates and Numerical Expressions in Arabic.	
Suitable for students who are studying in the following areas:	Computer Science and/or Computational Linguistics.	
Project Description:	<p>Dates and Numerical Expressions are challenging in Natural Language Processing because they represent an infinite set of possible expressions. In addition, they are represented in multiple script forms: digits , multi-word sequences (that can be inflected) or a mix of both.</p> <p>In parsing free text, detecting dates and numerical expressions as a multi-word sequence is crucial. A pre-processing step is needed to determine these cases and convert them into a normalized digit-based form.</p> <p>The internship aims to improve parsing systems by means of statistical methods or machine learning tools/techniques for the tasks of dates and number identification and normalization in Arabic free texts.</p>	
The Role of the student & benefits gained from participation in this project:¹	<ul style="list-style-type: none"> •Learn about research team environment and work, •Practice communication skills, •Develop research and programming skills, •Publication might be expected 	
Who will be working with you?	Dr. Lamia Tounsi Dr. Mohammed Attia	
Short description of the group:	<p>The National Centre for Language Technology is based in the School of Computing in Dublin city University. The Centre carries out basic and applied research in the areas of machine translation, natural language parsing, grammar induction, question answering, sentiment analysis, computer-aided language learning, software localisation, speech recognition and speech synthesis. Its researchers are drawn from the School of Computing, the School of Applied Languages and Intercultural Studies and the school of Electronic Engineering. The Centre is affiliated with the Center for Next Generation Localisation.</p> <p>http://www.nclt.dcu.ie/</p>	
Recommended Reading Material:	Habash, Nizar and Ryan Roth. <u>Identification of Naturally Occurring Numerical Expressions in Arabic</u> . In <i>Proceedings of the Language Resources and Evaluation Conference (LREC)</i> , Marrakech, Morocco, 2008.	
Other information:	The duration of the project: 8 weeks	
For further details on this project please contact:	Name: Phone: E-Mail: Website:	Dr. Lamia Tounsi 00 3535 (0)1 700 6905 lamia.tounsi@computing.dcu.ie

¹ This is an initial description of the role of the student and it is liable to change following discussions between the investigators and the student.

