CLTW 2014

The First Celtic Language Technology Workshop

Proceedings of the Workshop

A Workshop of the 25th International Conference on
Computational Linguistics (COLING 2014) August 23, 2014
Dublin, Ireland

ISBN 978-1-873769-32-4
Proceedings of the Celtic Language Technology Workshop (CLTW)
John Judge, Teresa Lynn, Monica Ward and Brian Ó Raghallaigh (eds.)

Introduction

Language Technology and Computational Linguistics research innovations in recent years have given us a great deal of modern language processing tools and resources for many languages. Basic language tools like spell and grammar checkers through to interactive systems like Siri, as well as resources like the Trillion Word Corpus, all fit together to produce products and services which enhance our daily lives.

Until relatively recently, languages with smaller numbers of speakers have largely not benefited from attention in this field. However, modern techniques in the field are making it easier to create language tools and resources from fewer resources in a faster time. In this light, many lesser spoken languages are making their way into the digital age through the provision of language technologies and resources.

The Celtic Language Technology Workshop (CLTW) series of workshops provides a forum for researchers interested in developing NLP (Natural Language Processing) resources and technologies for Celtic languages. As Celtic languages are under-resourced, our goal is to encourage collaboration and communication between researchers working on language technologies and resources for Celtic languages.

Welcome to the First Celtic Language Technology Workshop. We received 15 submissions, and after a rigorous review process, accepted 12 papers. Eight of which will be presented as oral presentations and 4 of which will be presented at the poster session.

Organising Committee:

John Judge, Centre for Global Intelligent Content (CNGL), Dublin City University

Teresa Lynn, Centre for Global Intelligent Content (CNGL), Dublin City University

Monica Ward, National Centre for Language Technology (NCLT), Dublin City University

Brian Ó Raghallaigh, Fiontar, Dublin City University

Program Committee:

Steven Bird, University of Melbourne, Australia
Aoife Cahill, Educational Testing Service (ETS), USA
Andrew Carnie, University of Arizona, USA
Jeremy Evas, Cardiff University, Wales
Mikel Forcada, Universitat d'Alacant, Spain
John Judge, CNGL, Dublin City University, Ireland
Teresa Lynn, CNGL, Dublin City University, Ireland
Ruth Lysaght, Université de Bretagne Occidentale, France
Neasa Ní Chiaráin, Trinity College Dublin, Ireland
Brian Ó Raghallaigh, Fiontar, Dublin City University, Ireland
Delyth Prys, Bangor University, Wales
Kevin Scannell, St. Louis University, USA
Mark Steedman, University of Edinburgh, Scotland
Nancy Stenson, University College Dublin, Ireland
Francis Tyers, Universitetet i Tromso, Norway
Elaine Uí Dhonnchadha, Trinity College Dublin, Ireland
Monica Ward, NCLT, Dublin City University, Ireland
Pauline Welby, CNRS, Université de Provence, France

Invited Speakers:

Kevin Scannell, St. Louis University, USA
Elaine Uí Dhonnchadha, Trinity College Dublin, Ireland

Sponsor:

Transpiral http://www.transpiral.com

Table of Contents

Conference Programme

23rd August 2014

+	Opening
09.00-09.30	Invited Talk by Elaine Uí Dhonnchadha
	(09.30-10.00) Morning Session 1
09.30–09.50	*Developing an Automatic Part-of-Speech Tagger for Scottish Gaelic* William Lamb and Samuel Danso
09.50–10.10	*Using Irish NLP resources in Primary School Education* Monica Ward
10.10–10.30	*Tools facilitating better use of online dictionaries: Technical aspects of Multidict, Wordlink and Clilstore* Caoimhin O Donnaile
	(10.30-11.00) Break
	(11.00-12.30) Morning Session 2
11.00–11.20	*Processing Mutations in Breton with Finite-State Transducers* Thierry Poibeau
11.20–11.40	*Statistical models for text normalization and machine translation* Kevin Scannell
11.40–12.00	*Cross-lingual Transfer Parsing for Low-Resourced Languages: An Irish Case Study* Teresa Lynn, Jennifer Foster, Mark Dras and Lamia Tounsi
12.00-12.30	TBA

23rd August 2014 (continued)

(12.30-14.00) Lunch

14.00-14.30 Invited Talk by Kevin Scannell

(14.30-15.30) Afternoon Session

14.30–14.50 *Irish National Morphology Database: a high-accuracy open-source dataset of Irish words*
Michal Boleslav Měchura

14.50–15.10 *Developing further speech recognition resources for Welsh*
Sarah Cooper, Dewi Jones and Delyth Prys

(15.10-15.30) Poster Boasters

(15.30-16.00) Break

(16.00-17.00) Poster/Networking Session

gdbank: The beginnings of a corpus of dependency structures and type-logical grammar in Scottish Gaelic
Colin Batchelor

Developing high-end reusable tools and resources for Irish-language terminology, lexicography, onomastics (toponymy), folkloristics, and more, using modern web and database technologies
Brian Ó Raghallaigh and Michal Boleslav Měchura

DECHE and the Welsh National Corpus Portal
Delyth Prys, Dewi Jones and Mared Roberts

Subsegmental language detection in Celtic language text
Akshay Minocha and Francis Tyers

17.00-17.05 Closing