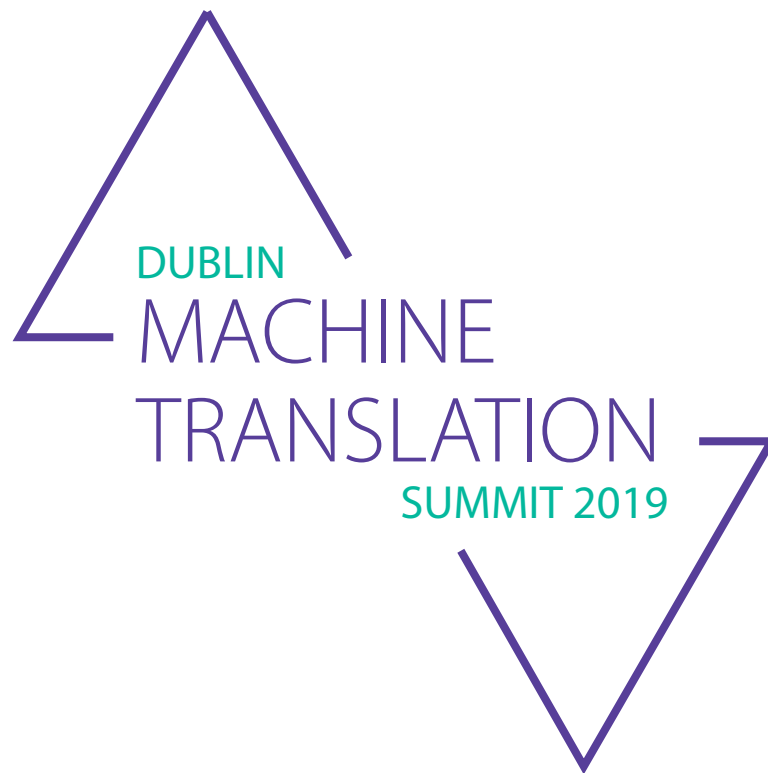


Machine Translation Summit XVII



Proceedings of the Celtic Language Technology Workshop 2019

<http://cl.indiana.edu/cltw19/>

19th August, 2019
Dublin, Ireland

Proceedings of the Celtic Language Technology Workshop 2019

<http://cl.indiana.edu/cltw19/>

19th August, 2019
Dublin, Ireland



© 2019 The authors. These articles are licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

Preface from the co-chairs of the workshop

These proceedings include the program and papers presented at the third Celtic Language Technology Workshop held in conjunction with the Machine Translation Summit in Dublin in August 2019. Celtic Languages are spoken in regions extending from the British Isles to the western seaboard of continental Europe, and communities in Argentina and Canada. They include Irish, Scottish Gaelic, Breton, Manx, Welsh and Cornish. Irish is an official EU language, both Irish and Welsh have official language status in their respective countries, and Scottish Gaelic recently saw an improvement in its prospects through the Gaelic Language (Scotland) Act 2005. The majority of others have co-official status, except Breton which has not yet received recognition as an official or regional language.

Until recently, the Celtic languages have lagged behind in the area of natural language processing (NLP) and machine translation (MT). Consequently, language technology research and resource provision for this language group was poor. In recent years, as the resource provision for minority and under-resourced languages has been improving, it has been extended and applied to some Celtic languages.. The CLTW community and workshop, inaugurated at COLING (Dublin) in 2014, provides such a forum for researchers interested in language/speech processing technologies and related resources for low-resourced languages, specifically those working with the Celtic languages.

The 11 accepted papers cover an extremely wide range of topics, including machine translation, tree-banking, computer-aided language learning, NLP of social media, information retrieval, Old Irish, the Welsh spoken in Argentina and named entity recognition.

We thank our invited speakers, Claudia Soria of the National Research Council of Italy and Kelly Davis of Mozilla. Claudia studies digital language diversity and Kelly is part of Mozilla's open-source Common Voice speech project. We also thank our authors and presenters for their hard work and workshop attendees for their participation, and of course we are very grateful to our programme committee for reviewing and providing invaluable feedback on the work published. Finally we thank Mozilla and the Department of Culture, Heritage and the Gaeltacht for financial support.

Workshop organisers

Teresa Lynn, Dublin City University

Delyth Prys, University of Bangor

Colin Batchelor, Royal Society of Chemistry

Francis M. Tyers, Indiana University and Higher School of Economics

Organisers

Workshop chairs

Teresa Lynn
Delyth Prys
Colin Batchelor
Francis M. Tyers

Dublin City University
University of Bangor
Royal Society of Chemistry
Indiana University,
Higher School of Economics

Programme committee

Andrew Carnie
Annie Foret
Arantza Diaz de Ilarraza
Brian Davis
Brian Ó Raghallaigh
Elaine Uí Dhonnchadha
Jeremy Evas
John Judge
John McCrae
Kepa Sarasola
Kevin Scannell
Mikel L. Forcada
Monica Ward
Montse Maritxalar
Nancy Stenson
Pauline Welby
William Lamb

University of Arizona
Université Rennes 1
Euskal Herriko Unibertsitatea
Maynooth University
Fiontar/Dublin City University
Trinity College Dublin
Cardiff University and Welsh Government
ADAPT Centre, Dublin City University
INSIGHT Centre, University of Galway
Euskal Herriko Unibertsitatea
St. Louis University
Universitat d'Alacant
Dublin City University
Euskal Herriko Unibertsitatea
University College Dublin
CNRS/University College Dublin
University of Edinburgh

Contents

Unsupervised multi-word term recognition in Welsh	1
<i>Irena Spasić, David Owen, Dawn Knight, Andreas Artemiou</i>	
Universal dependencies for Scottish Gaelic: syntax	7
<i>Colin Batchelor</i>	
Speech technology and Argentinean Welsh	16
<i>Elise Bell</i>	
Development of a Universal Dependencies treebank for Welsh	21
<i>Johannes Heinecke, Francis M. Tyers</i>	
Code-switching in Irish tweets: A preliminary analysis	32
<i>Teresa Lynn, Kevin Scannell</i>	
Embedding English to Welsh MT in a Private Company	41
<i>Myfyr Prys, Dewi Bryn Jones</i>	
Adapting Term Recognition to an Under-Resourced Language: the Case of Irish	48
<i>John P. McCrae, Adrian Doyle</i>	
Leveraging backtranslation to improve machine translation for Gaelic languages	58
<i>Meghan Dowling, Teresa Lynn, Andy Way</i>	
Improving full-text search results on <i>dúchas.ie</i> using language technology	63
<i>Brian Ó Raghallaigh, Kevin Scannell, Meghan Dowling</i>	
A Character-Level LSTM Network Model for Tokenizing the Old Irish text of the Würzburg Glosses on the Pauline Epistles	70
<i>Adrian Doyle, John P. McCrae, Clodagh Downey</i>	
A Green Approach for an Irish App (Refactor, reuse and keeping it real)	80
<i>Monica Ward, Maxim Mozgovoy, Marina Purgina</i>	