LREC 2020
Language Resources and Evaluation Conference
11–16 May 2020

Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)

PROCEEDINGS

Editors: Kathy McKeown (Chair), Douglas W. Oard, Elizabeth (Liz) Boschee, and Richard Schwartz.

Proceedings of the LREC 2020 Cross-Language Search and Summarization of Text and Speech (CLSSTS2020)

Edited by: Kathy McKeown (Chair), Douglas W. Oard, Elizabeth (Liz) Boschee, and Richard Schwartz.

ISBN: 979-10-95546-55-9
EAN: 9791095546559

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

Introduction

In today's global world, people may need access to information that only appears online in a language they do not speak. Cross-Language Information Retrieval (CLIR) enables end users to issue queries in their own language, but provides results from multiple languages around the world, often using translation so that the end user can quickly understand whether the retrieved results are relevant. Cross-language summarization can make it easier for an end user to determine if a document is relevant by providing a summary in the user's language of the foreign language document, highlighting the evidence for relevance. Alternatively, a summary can be used to get a sense of document meaning, when the document is not in the user's language. When the foreign language is a low-resource language, cross-language search and summarization are more difficult; translation capabilities may be poor and the lack of resources makes it difficult to train CLIR and summarization systems. To complicate matters even more, when the collection contains speech as well as text, producing accurate search results and generating comprehensible summaries is even more difficult.

This workshop aims to stimulate the collection and provision of resources that can improve systems that perform cross-language search and summarization. Papers were solicited that describe recent and current research in these areas, that describe relevant resources, or that stake out positions on the directions in which the authors think the field should move.

Had the workshop proceeded in person, it would have featured a keynote speech by Carl Rubino, program manager of the IARPA MATERIAL program (USA). Carl was planning to describe the program and the languages studied, as well as metrics that the program uses during its evaluations, paying particular attention to the correlation between linguistic properties and system performance. We had also planned a second keynote speech by Julio Gonzalo, Universidad Nacional de Educación a Distancia (Madrid, Spain), who has recently been working on reputation reports, summaries of what is being said about an entity with a focus on reputational consequences. They have collected a large multilingual test collection for the reputation monitoring problem, with over half a million manual annotations for several tasks on twitter data, including named-entity disambiguation, reputational polarity, topic detection, reputational alerts, reputation reports, opinion maker identification, reputational dimensions, and author profiling.

To set the stage, the organizers provide two small spoken language test collections that include waveforms, transcriptions and possibly queries with relevance judgments. These are conversational genres, one in Somali (a very-low resource language) and the other in Bulgarian (a moderate-resource language) both of which include approximately 80 hours of speech.

Organizing Committee

Kathy McKeown, Columbia University (USA), Chair
James Allan, University of Massachusetts at Amherst (USA)
Lu Wang, Northeastern University (USA)
Douglas W. Oard, University of Maryland (USA)
Steve Renals, University of Edinburgh (UK)
Elizabeth (Liz) Boschee, USC/Information Sciences Institute (USA)
Richard Schwartz, Raytheon BBN Technologies (USA), Editor

Program Committee

Eneko Agirre, University of the Basque Country (Spain)
Piyush Arora, American Express Big Data Labs (India)
Mohit Bansal, University of North Carolina (USA)
Nicola Ferro, University of Padua (Italy)
Petra Galuscakova, University of Maryland (USA)
Jan Hajic, Charles University (Czech Republic)
Gareth Jones, Dublin City University (Ireland)
Damianos Karakos, Reytheon BBN Technologies, (USA)
Jonathan May, University of Southern California Information Sciences Institute (USA)
Jessica Ouyang, University of Texas at Dallas (USA)
Pavel Pecina, Charles University (Czech Republic)
Kay Peterson, NIST (USA)
Dragomir Radev, Yale University (USA)
Hussein Suleman, University of Cape Town (South Africa)
Audrey Tong, NIST (USA)
Xabier Saralegi Urizar, Elhuyar Foundation (Spain)
Ilya Zavorin, Bluemont Technology (USA)
Rui Zhang, Yale University (USA)

Table of Contents

The Effect of Linguistic Parameters in CLIR Performance
Carl Rubino . . . 1

Corpora for Cross-Language Information Retrieval in Six Less-Resourced Languages
Ilya Zavorin, Aric Bills, Cassian Corey, Michelle Morrison, Audrey Tong and Richard Tong . . . 7

MATERIALizing Cross-Language Information Retrieval: A Snapshot
Petra Galuscakova, Douglas Oard, Joe Barrow, Suraj Nair, Shing Han-Chin, Elena Zotkina, Ramy Eskander and Rui Zhang . . . 14

SEARCHER: Shared Embedding Architecture for Effective Retrieval
Joel Barry, Elizabeth Boschee, Marjorie Freedman and Scott Miller . . . 22

Cross-lingual Information Retrieval with BERT
Zhuolin Jiang, Amro El-Jaroudi, William Hartmann, Damianos Karakos and Lingjun Zhao . . . 26

A Comparison of Unsupervised Methods for Ad hoc Cross-Lingual Document Retrieval
Elaine Zosa, Mark Granroth-Wilding and Lidia Pivovarova . . . 32

Reformulating Information Retrieval from Speech and Text as a Detection Problem
Damianos Karakos, Rabih Zbib, William Hartmann, Richard Schwartz and John Makhoul . . . 38

The 2019 BBN Cross-lingual Information Retrieval System
Le Zhang, Damianos Karakos, William Hartmann, Manaj Srivastava, Lee Tarlin, David Akodes, Sanjay Krishna Gouda, Numra Bathool, Lingjun Zhao, Zhuolin Jiang, Richard Schwartz and John Makhoul . . . 44

What Set of Documents to Present to an Analyst?
Richard Schwartz, John Makhoul, Lee Tarlin and Damianos Karakos . . . 52

An Investigative Study of Multi-Modal Cross-Lingual Retrieval
Piyush Arora, Dimitar Shterionov, Yasufumi Moriya, Abhishek Kaushik, Daria Dzendzik and Gareth Jones . . . 58

Subtitles to Segmentation: Improving Low-Resource Speech-to-TextTranslation Pipelines
David Wan, Zhengping Jiang, Chris Kedzie, Elsbeth Turcan, Peter Bell and Kathy McKeown . . . 68