Cupral 2017

The First Workshop on Curation and Applications of Parallel and Comparable Corpora

Proceedings of the Workshop

November 27, 2017
Taipei, Taiwan

ISBN 978-1-948087-05-6

Introduction

The First Workshop on Curation and Applications of Parallel and Comparable Corpora (Cupral 2017) took place on Monday, November 27, 2017 in Taipei, Taiwan, immediately preceding the International Conference on Natural Language Processing (IJCNLP 2017).

The focus of our workshop was to explore the multifarious aspects of effective document alignment in multimodal and multilingual context. Most businesses operating across international borders understand the value of localization. In order to make a connection they have to be able to speak the language of their customers. Websites, marketing materials, news and other high-impact elements should all be thoroughly localized, which can mean a combination of computer vision (CV) and text processing in many target languages.

Clearly, techniques of Natural Language Processing (NLP) and Information Retrieval (IR) can be incredibly useful and further their combination with CV can potentially improve state-of-the-art document alignment research. Additionally, the aligned multimodal documents can be seamlessly used to improve the quality of predictive analytics on multi-modal data involving both text and images, e.g. the associated images of news articles may be utilized to help improve the ranks of these articles in a search engine or to translate the article better in a different language.

The workshop aimed to provide a forum for researchers working on related fields to present their results and insights. Our goal was to bring together researchers from diverse fields, such as CV, IR and NLP, who can potentially contribute to improving the quality of multimodal document alignment and its utilization in research and industrial data analytics tasks. The workshop was a starting point for an international platform dedicated to new method and techniques on aligning multimodal and multilingual documents, and exploring the use of such technology in NLP or IR.

Manoj Kumar Chinnakotla from Microsoft gave the invited on “Leveraging Parallel and Comparable Corpora for Multilingual NLP”.

We would like to thank the members of the Program Committee for their timely reviews. We also would like to thank the participants of the workshop for the interesting discussions around our Cupal 2017 topics.

Haithem Afli & Chao-Hong Liu

Organizers:

Haithem Afli, ADAPT Centre, Dublin City University
Chao-Hong Liu, ADAPT Centre, Dublin City University

Program Committee:

Debasis Ganguly, Dublin Research Lab, IBM Ireland
Longyue Wang, Dublin City University
Alberto Poncelas, Dublin City University
Iacer Calixto, ADAPT Centre, Dublin City University

Additional Reviewers:

Walid Aransa, LIUM, Le Mans University
Pintu Lohar, Dublin City University

Invited Speaker:

Manoj Kumar Chinnakotla, Microsoft, India

Table of Contents

Workshop Program

Monday, November 27, 2017

08:00–09:10 ***Registration***

09:15–9:30 ***Opening Remarks***

09:30–10:30 ***Keynote Talk by Manoj Kumar Chinnakotla (Microsoft) on "Leveraging Parallel and Comparable Corpora for Multilingual NLP"***

10:30–11:00 ***Coffee Break***

Presentations Session

11:00–11:30 *Building a Better Bitext for Structurally Different Languages through Self-training*
Jungyeul Park, Loic Dugast, Jeen-Pyo Hong, Chang-Uk Shin and Jeong-Won Cha

11:30–12:00 *MultiNews: A Web collection of an Aligned Multimodal and Multilingual Corpus*
Haithem Afli, Pintu Lohar and Andy Way

12:00–12:30 *Learning Phrase Embeddings from Paraphrases with GRUs*
zhihao zhou, Lifu Huang and Heng Ji

12:30–12:45 ***Closing Session***