EMNLP 2019

Aggregating and Analysing Crowdsourced Annotations for NLP

Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP (AnnoNLP)

November 3rd, 2019
Hong Kong, China

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-950737-80-2

Introduction

Welcome to the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP. We received 16 submissions and we accepted 7 of them. We are excited to also include two invited talks and one spotlight presentation.

Crowdsourcing, whether through microwork platforms or through Games with a Purpose, is increasingly used as an alternative to traditional expert annotation, achieving comparable annotation quality at lower cost and offering greater scalability. The NLP community has enthusiastically adopted crowdsourcing to support work in tasks such as coreference resolution, sentiment analysis, textual entailment, named entity recognition, word similarity, word sense disambiguation, and many others. This interest has also resulted in the organization of a number of workshops at ACL and elsewhere, from as early as "The People's Web meets NLP" in 2009. These days, general purpose research on crowdsourcing can be presented at HCOMP or CrowdML, but the need for workshops more focused on the use of crowdsourcing in NLP remains. In particular, NLP-specific methods are typically required for the task of aggregating the interpretations provided by the annotators.

Most existing work on aggregation methods is based on a common set of assumptions: 1) independence between the true classes, 2) the set of classes the coders can choose from is fixed across the annotated items, and 3) there is one true class per item. However, for many NLP tasks such assumptions are not entirely appropriate. For example, sequence labelling tasks (e.g., NER, tagging) have an implicit inter-label dependence. In other tasks such as coreference the labels the coders can choose from are not fixed but depend on the mentions from each document. Furthermore, in many NLP tasks, the data items can have more than one interpretation. Such cases of ambiguity also affect the reliability of existing gold standard datasets (often labelled with a single interpretation even though expert disagreement is a well-known issue). This former point motivates the research on alternative, complementary evaluation methods, but also the development of multi-label datasets.

The workshop aims to bring together researchers interested in methods for aggregating and analysing crowdsourced data for NLP-specific tasks which relax the aforementioned assumptions. We also invited work on ambiguous, subjective or complex annotation tasks which received less attention in the literature.

We would like to thank the program committee, all authors and invited speakers, and hope you enjoy the workshop.

Silviu Paun and Dirk Hovy
November 2019

Organizers:

Silviu Paun, Queen Mary University of London
Dirk Hovy, Bocconi University

Program Committee:

Beata Beigman Klebanov, Princeton (USA)
Bob Carpenter, Columbia University (USA)
Jon Chamberlain, University of Essex (UK)
Anca Dumitrache, Vrije Universiteit Amsterdam (Netherlands)
Paul Felt, IBM (USA)
Udo Kruschwitz, University of Essex (UK)
Matthew Lease, University of Texas at Austin (USA)
Massimo Poesio, Queen Mary University of London (UK)
Edwin Simpson, Technische Universität Darmstadt (Germany)
Henning Wachsmuth, Universität Paderborn (Germany)

Additional Reviewers:

Chris Madge, Queen Mary University of London (UK)
Juntao Yu, Queen Mary University of London (UK)

Invited Speaker:

Jordan Boyd-Graber, University of Maryland (USA)
Edwin Simpson, Technische Universität Darmstadt (Germany)

Table of Contents

Conference Program

Sunday, November 3, 2019

9:00–10:30 Session 1

09:00–09:10 ***Welcome remarks***

09:10–10:10 *Invited Talk*
Jordan Boyd-Graber, University of Maryland

10:10–10:30 *Dependency Tree Annotation with Mechanical Turk*
Stephen Tratz

10:30–11:00 Coffee Break

11:00–12:20 Session 2

11:00–11:30 *Word Familiarity Rate Estimation Using a Bayesian Linear Mixed Model*
Masayuki Asahara

11:30–12:00 *Leveraging syntactic parsing to improve event annotation matching*
Camiel Colruyt, Orphée De Clercq and Véronique Hoste

12:00–12:20 *A Dataset of Crowdsourced Word Sequences: Collections and Answer Aggregation for Ground Truth Creation*
Jiyi Li and Fumiyo Fukumoto

Sunday, November 3, 2019 (continued)

12:20–14:00 Lunch break

14:00–15:20 Session 3

14:00–15:00 *Invited Talk*
Edwin Simpson, Technische Universität Darmstadt

15:00–15:20 *Distance-based Consensus Modeling for Complex Annotations*
Alexander Braylan and Matthew Lease

15:20–16:00 Afternoon coffee break

16:00–17:20 Session 4

16:00–16:30 *Crowd-sourcing annotation of complex NLU tasks: A case study of argumentative content annotation*
Tamar Lavee, Lili Kotlerman, Matan Orbach, Yonatan Bilu, Michal Jacovi, Ranit Aharonov and Noam Slonim

16:30–17:00 *Computer Assisted Annotation of Tension Development in TED Talks through Crowdsourcing*
Seungwon Yoon, Wonsuk Yang and Jong Park

17:00–17:20 *CoSSAT: Code-Switched Speech Annotation Tool*
Sanket Shah, Pratik Joshi, Sebastin Santy and Sunayana Sitaram