Lecture Notes in Computer Science 13390

More information about this series at https://link.springer.com/bookseries/558

Alberto Barrón-Cedeño ·
Giovanni Da San Martino · Mirko Degli Esposti ·
Fabrizio Sebastiani · Craig Macdonald ·
Gabriella Pasi · Allan Hanbury · Martin Potthast ·
Guglielmo Faggioli · Nicola Ferro (Eds.)

Experimental IR Meets Multilinguality, Multimodality, and Interaction

13th International Conference of the CLEF Association, CLEF 2022
Bologna, Italy, September 5–8, 2022
Proceedings

Springer

Editors
Alberto Barrón-Cedeño
University of Bologna
Forlì, Italy

Mirko Degli Esposti
University of Bologna
Bologna, Italy

Craig Macdonald
University of Glasgow
Glasgow, UK

Allan Hanbury
TU Wien
Vienna, Austria

Guglielmo Faggioli
University of Padua
Padova, Italy

Giovanni Da San Martino
University of Padua
Padova, Italy

Fabrizio Sebastiani
Instituto di Scienza e Tecnologie dell'
Informazione "Alessandro Faedo"
Pisa, Italy

Gabriella Pasi
University Milano-Bicocca
Milan, Italy

Martin Potthast
Leipzig University
Leipzig, Germany

Nicola Ferro
University of Padua
Padova, Italy

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-031-13642-9 ISBN 978-3-031-13643-6 (eBook)
https://doi.org/10.1007/978-3-031-13643-6

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Since 2000, the Conference and Labs of the Evaluation Forum (CLEF) has played a leading role in stimulating research and innovation in the domain of multimodal and multilingual information access. Initially founded as the Cross-Language Evaluation Forum and running in conjunction with the European Conference on Digital Libraries (ECDL/TPDL), CLEF became a standalone event in 2010 combining a peer-reviewed conference with a multi-track evaluation forum. The combination of the scientific program and the track-based evaluations at the CLEF conference creates a unique platform to explore information access from different perspectives, in any modality and language.

The CLEF conference has a clear focus on experimental information retrieval (IR) as seen in evaluation forums (like the CLEF Labs, TREC, NTCIR, FIRE, MediaEval, RomIP, TAC) with special attention paid to the challenges of multimodality, multilinguality, and interactive search, ranging from unstructured to semi-structured and structured data. The CLEF conference invites submissions on new insights demonstrated by the use of innovative IR evaluation tasks or in the analysis of IR test collections and evaluation measures, as well as on concrete proposals to push the boundaries of the Cranfield/TREC/CLEF paradigm.

CLEF 2022[1] was organized by the University of Bologna, Italy, and held during September 5–8, 2022. Despite the continued outbreak of the COVID-19 pandemic, the improvement of the overall situation allowed for organizing CLEF 2022 as an in-person event, after two editions—CLEF 2020 and 2021—were forced to be virtual only. The conference format remained the same as in past years and consisted of keynotes, contributed papers, lab sessions, and poster sessions, including reports from other benchmarking initiatives from around the world. All sessions were held in person but also allowed for remote participation for those who were not able to attend physically.

CLEF 2022 continued the initiative introduced in the 2019 edition, during which the European Conference for Information Retrieval (ECIR) and CLEF joined forces: ECIR 2022 hosted a special session dedicated to CLEF Labs where lab organizers presented the major outcomes of their labs and their plans for ongoing activities, which was followed by a poster session to encourage discussion during the conference. This was reflected in the ECIR 2022 proceedings, where CLEF Lab activities and results were reported as short papers. The goal was not only to engage the ECIR community in CLEF activities but also to disseminate the research results achieved during CLEF evaluation cycles through the submission of papers to ECIR.

The following scholars were invited to give a keynote talk at CLEF 2022: Rita Cucchiara (University of Modena and Reggio Emilia, Italy) and Benno Stein (Bauhaus-Universität Weimar, Germany).

CLEF 2022 received a total of 14 scientific submissions, of which a total of 10 papers (seven long and three short) were accepted. Each submission was reviewed by three Program Committee (PC) members, and the program chairs oversaw the reviewing

[1] https://clef2022.clef-initiative.eu/.

and follow-up discussions. Ten countries are represented in the accepted papers, as several of them were a product of international collaboration. This year, researchers addressed the following important challenges in the community: authorship attribution, fake news detection and news tracking, noise-detection in automatically transferred relevance judgments, the impact of online education on children's conversational search behavior, analysis of multi-modal social media content, knowledge graphs for sensitivity identification, a fusion of deep learning and logic rules for sentiment analysis, medical concept normalization, and domain-specific information extraction.

Like in previous editions, since 2015, CLEF 2022 invited CLEF lab organizers to nominate a "best of the labs" paper that was reviewed as a full paper submission to the CLEF 2022 conference, according to the same review criteria and PC. In total, seven full papers were accepted for this "best of the labs" section.

The conference integrated a series of workshops presenting the results of lab-based comparative evaluations. CLEF 2022 was the 13th edition of the CLEF conference and it was the 23rd year of the CLEF initiative as a forum for IR evaluation. A total of 15 lab proposals were received and evaluated in peer review based on their innovation potential and the quality of the resources created. The 14 selected labs represented scientific challenges based on new datasets and real-world problems in multimodal and multilingual information access. These datasets provide unique opportunities for scientists to explore collections, to develop solutions for these problems, to receive feedback on the performance of their solutions, and to discuss the challenges with peers at the workshops. In addition to these workshops, the labs reported results of their year long activities in overview talks and lab sessions. Overview papers describing each of the labs are provided in this volume. The full details for each lab are contained in a separate publication, the Working Notes[2].

The 14 labs running as part of CLEF 2022 comprised mainly labs that continued from previous editions at CLEF (ARQMath, BioASQ, CheckThat!, CheMU, eRisk, ImageCLEF, LifeCLEF, PAN, SimpleText, and Touché) and newer pilot/workshop activities (HIPE, iDPP, JOKER, and LeQUA). In the following we give a few details for each of the labs organized at CLEF 2022 (presented in alphabetical order):

ARQMath: Answer Retrieval for Mathematical Questions[3] aims to advance math-aware search and the semantic analysis of mathematical notation and texts. It offered the following tasks. Task 1: Answer Retrieval, given a math question post, return relevant answer posts. Task 2: Formula Retrieval; given a formula in a math question post, return relevant formulas from both question and answer posts. Task 3: Open Domain Question Answering, given a math question post, return an automatically generated answer that is comprised of excerpts from arbitrary sources and/or machine generated.

BioASQ: Large-scale Biomedical Semantic Indexing and Answering[4] aims to push the research frontier towards systems that use the diverse and voluminous information available online to respond directly to the information needs of biomedical scientists.

[2] Faggioli, G., Ferro, N., Hanbury, A., and Potthast, M. (Eds.). (2022). *CLEF 2022 Working Notes*. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073.

[3] https://www.cs.rit.edu/~dprl/ARQMath.

[4] http://www.bioasq.org/workshop2022.

It offered the following tasks. Task 1: Large-scale Online Biomedical Semantic Indexing involved classifying new PubMed documents before PubMed curators annotate (in effect, classify) them manually into classes from the MeSH hierarchy. Task 2: Biomedical Semantic Question Answering used benchmark datasets of biomedical questions, in English, along with gold standard (reference) answers constructed by a team of biomedical experts. The participants had to respond with relevant articles and snippets from designated resources, as well as exact and "ideal" answers. Task 3: DisTEMIST - Disease Text Mining and Indexing Shared Task focused on the recognition and indexing of diseases in medical documents in Spanish by posing subtasks on (1) indexing medical documents with controlled terminologies; (2) automatic detection indexing textual evidence (i.e., disease entity mentions in text); and (3) normalization of these disease entity mentions to terminologies. Task 4: Task Synergy - Question Answering for Developing Problems, biomedical experts posed unanswered questions for the developing problem of COVID-19, received the responses provided by the participating systems, and provided feedback, together with updated questions, in an iterative procedure that aimed to facilitate the incremental understanding of COVID-19.

CheckThat!: Lab on Fighting the COVID-19 Infodemic and Fake News Detection[5] aims at fighting misinformation and disinformation in social media, in political debates, and in the news. It offered the following tasks. Task 1: Identifying Relevant Claims in Tweets focused on disinformation related to the ongoing COVID-19 infodemic politics. It asked participants to identify which posts in a Twitter stream are worth fact-checking, contain a verifiable factual claim, are harmful to society, and why. This task was offered in Arabic, Bulgarian, Dutch, English, Spanish, and Turkish. In Task 2: Detecting Previously Fact-Checked Claims, given a check-worthy claim, and a set of previously-checked claims, participants had to determine whether the claim has been previously fact-checked with respect to a collection of fact-checked claims. The text could be a tweet or a sentence from a political debate. The task was offered in Arabic and English. In Task 3: Fake News Detection, given the text and the title of a news article, participants had to determine whether the main claim made in the article is true, partially true, false, or other (e.g., articles in dispute and unproven articles). This task was offered in English and German.

ChEMU: Cheminformatics Elsevier Melbourne University[6] focuses on information extraction in chemical patents, including tasks ranging from document- to expression-level. It offered the following tasks. Task 1a: Named Entity Recognition was aimed at identifying chemical compounds, their specific types, temperatures, reaction times, yields, and the label of the reaction. Task 1b: Event Extraction acknowledged that a chemical reaction leading to an end product often consists of a sequence of individual event steps. The task was to identify those steps involving the chemical entities recognized from Task 1a. Task 1c: Anaphora Resolution required requires the resolution of anaphoric dependencies between expressions in chemical patents. The participants were required to find five types of anaphoric relationships in chemical patents: coreference, reaction-associated, work-up, contained, and transform. In Task 2a: Chemical Reaction Reference Resolution, given a reaction description, participants had to identify

[5] https://sites.google.com/view/clef2022-checkthat.
[6] http://chemu2022.eng.unimelb.edu.au/.

references to other reactions that the reaction relates to, and to the general conditions that it depends on. Task 2b: Table Semantic Classification involved classifying tables in chemical patents into eight categories based on their contents.

eRisk: Early Risk Prediction on the Internet[7] explores the evaluation methodology, effectiveness metrics, and practical applications (particularly those related to health and safety) of early risk detection on the Internet. The main goal is to pioneer a new interdisciplinary research area that would be potentially applicable to a wide variety of situations and to many different personal profiles. Examples include potential pedophiles, stalkers, individuals that could fall into the hands of criminal organizations, people with suicidal inclinations, or people susceptible to depression. It offered the following tasks. Task 1: Early Detection of Signs of Pathological Gambling consisted of sequentially processing pieces of evidence and detecting early traces of pathological gambling (also known as compulsive gambling or disordered gambling), as soon as possible. Task 2: Early Detection of Depression consisted of sequentially processing pieces of evidence and detecting early traces of depression as soon as possible. Tasks 1 and 2 were mainly concerned with evaluating text mining solutions and thus concentrated on texts written in social media. Task 3: Measuring the Severity of the Signs of Eating Disorders consisted of estimating the level of features associated with a diagnosis of eating disorders from a thread of user submissions. For each user, the participants were given a history of postings and had to complete a standard eating disorder questionnaire (based on the evidence found in the history of postings).

HIPE: Named Entity Recognition and Linking in Multilingual Historical Documents[8] aims to assess and advance the development of robust, adaptable, and transferable named entity processing systems. Compared to the first HIPE edition in 2020, HIPE 2022 confronted systems with the challenges of dealing with more languages, learning domain-specific entities, and adapting to diverse annotation schemas. It offered the following tasks. Task 1: Named Entity Recognition and Classification (NERC) featured two subtasks—NERC-coarse on high-level entity types, for all languages, and NERC-fine on finer-grained entity types, for English, French, and German only. Task 2: Named Entity Linking (EL) involved the linking of named entity mentions to a unique referent in a knowledge base (Wikidata) or to a NIL node if the mention does not have a referent in the knowledge base.

iDPP: Intelligent Disease Progression Prediction[9] aims to design and develop an evaluation infrastructure for AI algorithms able to (1) better describe the mechanism of Amyotrophic Lateral Sclerosis (ALS) disease; (2) stratify patients according to their phenotype assessed throughout the disease evolution; and (3) predict ALS progression in a probabilistic, time dependent fashion. It offered the following tasks. Task 1: Ranking Risk of Impairment focused on ranking patients based on the risk of impairment in specific domains. We used the ALSFRS-R scale to monitor speech, swallowing, handwriting, dressing/hygiene, walking, and respiratory ability in time and asked participants to rank patients based on time to event risk of experiencing impairment in each specific domain. Task 2: Predicting Time of Impairment refined Task 1 by asking participants

[7] https://erisk.irlab.org/.

[8] https://hipe-eval.github.io/HIPE-2022/.

[9] https://brainteaser.health/open-evaluation-challenges/idpp-2022/.

to predict when specific impairments will occur (i.e., in the correct time window) by assessing model calibration in terms of the ability of the proposed algorithms to estimate a probability of an event close to the true probability within a specified time window. Task 3: Explainability of AI Algorithms called for position papers to start a discussion on AI explainability, including proposals on how the single patient data can be visualized in a multivariate fashion contextualizing its dynamic nature and the model predictions together with information on the predictive variables that most influence the prediction.

ImageCLEF: Multimedia Retrieval[10] promotes the evaluation of technologies for annotation, indexing, classification, and retrieval of multi-modal data, with the objective of providing information access to large collections of images in various usage scenarios and domains. It offered the following tasks. Task 1: ImageCLEFmedical focused on interpreting and summarizing the insights gained from radiology images, i.e. developing systems that are able to predict the UMLS concepts from visual image content, and implementing models to predict captions for given radiology images. The tuberculosis subtask called for systems that are able to detect lung cavern regions rather than simply provide a label for the CT images. Task 2: ImageCLEFcoral fostered tools for creating three-dimensional models of underwater coral environments. It required participants to label coral underwater images with types of benthic substrate, together with their bounding box, and to segment and parse each coral image into different image regions associated with benthic substrate types. Task 3: ImageCLEFaware concerned the online disclosure of personal data, which often has effects that go beyond the initial context in which data were shared. Participants were required to provide automatic rankings of photographic user profiles in a series of real-life situations, such as searching for a bank loan, accommodation, a waiter job, or a job in IT, with the ranking based on an automatic analysis of profile images and the aggregation of individual results. Task 4: ImageCLEFfusion involved system fusion—exploiting the complementary nature of individual systems to boost performance. Participants were tasked with creating novel ensembling methods that are able to significantly increase the performance of precomputed inducers in various use-case scenarios, such as visual interestingness and video memorability prediction.

JokeR: Automatic Wordplay and Humour Translation Workshop[11] aims to bring together translators and computer scientists to work on an evaluation framework for creative language, including data and metric development, and to foster work on automatic methods for wordplay translation. It offered the following tasks. Pilot Task 1: Classify and Interpret Wordplay involved, classifying single words containing wordplay according to a given typology, and providing lexical-semantic interpretations. Pilot Task 2: Translate Single Term Wordplay required the translation of single words containing wordplay. Pilot Task 3: Translate Phase Wordplay involved the translation of entire phrases that subsume or contain wordplay. In Task 4: Unshared Task, we welcomed submissions that use our data in other ways.

LeQua: Learning to Quantify[12] aims to allow the comparative evaluation of methods for "learning to quantify" in textual datasets; i.e. methods for training predictors of

[10] https://www.imageclef.org/2022.
[11] http://joker-project.com/.
[12] https://lequa2022.github.io/.

the relative frequencies of the classes of interest in sets of unlabeled textual documents. These predictors (called "quantifiers") are required to issue predictions for several such sets, some of them characterized by class frequencies radically different from the ones of the training set. This first edition of LeQua offered the following tasks. In Task 1 participants were provided with documents already converted into vector form; the task was thus suitable for participants who do not wish to engage in generating representations for the textual documents, but wanted instead to concentrate on optimizing the methods for learning to quantify. In Task 2 participants were provided with the raw text of the documents; the task was thus suitable for participants who also wished to engage in generating suitable representations for the textual documents, or to train end-to-end systems.

LifeCLEF: Biodiversity Identification and Prediction[13] aims to stimulate research in data science and machine learning for biodiversity monitoring. It offered the following tasks. Task 1: BirdCLEF involved bird species recognition in audio soundscapes. Task 2: PlantCLEF concerned image-based plant identification on a global scale (300K classes). Task 3: GeoLifeCLEF required location-based prediction of species based on environmental and occurrence data. Task 4: SnakeCLEF involved snake species identification in medically important scenarios. Task 5: FungiCLEF involved fungi recognition from image and metadata.

PAN: Digital Text Forensics and Stylometry[14] focuses on digital text forensics and stylometry, studying how to quantify writing style and improve authorship technology. It offered the following tasks. Task 1: Authorship Verification, given two texts, determine if they are written by the same author. Task 2: IROSTEREO, profiling Irony and Stereotype Spreaders on Twitter, given a Twitter feed, determine whether its author spreads Irony and Stereotypes. Task 3: Style Change Detection, given a document, determine the number of authors and at which positions the author changes. Task 4: Trigger Warning Prediction, given a document, determine whether its content warrants a warning of potential negative emotional responses in readers.

SimpleText: Automatic Simplification of Scientific Texts[15] addresses the challenges of text simplification approaches in the context of promoting scientific information access, by providing appropriate data and benchmarks, and creating a community of NLP and IR researchers working together to resolve one of the greatest challenges of today. It offered the following tasks. Task 1: What is in (or out)? Select passages to include in a simplified summary, given a query. Task 2: What is unclear? Given a passage and a query, rank terms/concepts that must be explained to understand this passage (definitions, context, applications, . . .). Task 3: Rewrite this! Given a query, simplify passages from scientific abstracts. Task 4: Unshared task, we welcomed any submission that uses our data.

Touché: Argument Retrieval[16] focuses on decision making processes, be it at the societal or at the personal level, that often come to a point where one side challenges the other with a why-question, which is a prompt to justify some stance based on arguments.

[13] https://www.imageclef.org/LifeCLEF2022.
[14] http://pan.webis.de/.
[15] http://simpletext-project.com/.
[16] https://touche.webis.de/.

Since technologies for argument mining are maturing at a rapid pace, ad-hoc argument retrieval has also become within reach. Touché offered the following tasks. Task 1: Argument Retrieval for Controversial Questions, given a controversial topic and a collection of argumentative documents, participants had to retrieve and rank sentences (the main claim and its most important premise in the document) that convey key points pertinent to the controversial topic. Task 2: Argument Retrieval for Comparative Questions, given a comparative topic and a collection of documents, participants had to retrieve relevant argumentative passages for either the compared object or for both objects and detect the respective stances with regard to the object in question. Task 3: Image Retrieval for Arguments, given a controversial topic, participants had to retrieve images (from web pages) for each stance (pro/con) that show support for that stance.

The success of CLEF 2022 would not have been possible without the huge effort of numerous people and organizations, including the CLEF Association[17], the Program Committee, the Lab Organizing Committee, the reviewers, and the many students and volunteers who contributed.

Finally, we thank the University of Bologna (with special mention to the DIT, DIFA, and DISI departments), the Department of Mathematics of the University of Padua, and the AI4media H2020 project for their invaluable support. We thank the Friends of SIGIR program for covering the registration fees for a number of student delegates.

July 2022

Alberto Barrón-Cedeño
Giovanni Da San Martino
Mirko Degli Esposti
Fabrizio Sebastiani
Craig Macdonald
Gabriella Pasi
Allan Hanbury
Martin Potthast
Guglielmo Faggioli
Nicola Ferro

[17] http://www.clef-initiative.eu/association.

Organization

CLEF 2022, the Conference and Labs of the Evaluation Forum – Experimental IR meets Multilinguality, Multimodality, and Interaction, was hosted by University of Bologna, Italy.

General Chairs

Alberto Barrón-Cedeño	University of Bologna, Italy
Giovanni Da San Martino	University of Padua, Italy
Mirko Degli Esposti	University of Bologna, Italy
Fabrizio Sebastiani	ISTI-CNR, Italy

Program Chairs

Craig Macdonald	University of Glasgow, UK
Gabriella Pasi	University of Milano-Bicocca, Italy

Lab Chairs

Allan Hanbury	Vienna University of Technology, Austria
Martin Potthast	Leipzig University, Germany

Lab Mentorship Chair

Paolo Rosso	Universitat Politécnica de Valencia, Spain

Proceedings Chairs

Guglielmo Faggioli	University of Padua, Italy
Nicola Ferro	University of Padua, Italy

Program Committee

Martin Braschler	ZHAW Zurich University of Applied Sciences, Switzerland
Annalina Caputo	Dublin City University, Ireland
Fabio Crestani	USI, Switzerland
Giorgio Maria Di Nunzio	University of Padua, Italy

Guglielmo Faggioli	University of Padua, Italy
Jussi Karlgren	Spotify, USA
Liadh Kelly	Maynooth University, Ireland
Maria Maistro	University of Copenhagen, Denmark
Stefano Marchesin	University of Padua, Italy
Yashar Moshfeghi	University of Strathclyde, UK
Franco Maria Nardini	ISTI-CNR, Italy
Jian-Yun Nie	University of Montreal, Canada
Iadh Ounis	University of Glasgow, UK
Paolo Rosso	Universitat Politècnica de València, Spain
Eric Sanjuan	Laboratoire Informatique d'Avignon, Université d'Avignon, France
Ting Su	University of Glasgow, UK
Hanna Suominen	ANU, Australia
Theodora Tsirika	CERTH-ITI, Greece
Marco Viviani	Università degli Studi di Milano-Bicocca, Italy
Ellen Voorhees	National Institute for Standards and Technology (NIST), USA
Christa Womser-Hacker	University of Hildesheim, Germany

CLEF Steering Committee

Steering Committee Chair

Nicola Ferro	University of Padua, Italy

Deputy Steering Committee Chair for the Conference

Paolo Rosso	Universitat Politècnica de València, Spain

Deputy Steering Committee Chair for the Evaluation Labs

Martin Braschler	Zurich University of Applied Sciences, Switzerland

Members

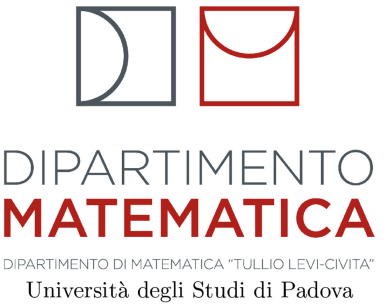
Khalid Choukri	Evaluations and Language resources Distribution Agency (ELDA), France
Fabio Crestani	Università della Svizzera italiana, Switzerland
Carsten Eickhoff	Brown University, USA
Norbert Fuhr	University of Duisburg-Essen, Germany

Lorraine Goeuriot	Université Grenoble Alpes, France
Julio Gonzalo	National Distance Education University (UNED), Spain
Donna Harman	National Institute for Standards and Technology (NIST), USA
Bogdan Ionescu	Politehnica University of Bucharest, Romania
Evangelos Kanoulas	University of Amsterdam, The Netherlands
Birger Larsen	University of Aalborg, Denmark
David E. Losada	Universidade de Santiago de Compostela, Spain
Mihai Lupu	Vienna University of Technology, Austria
Maria Maistro	University of Copenhagen, Denmark
Josiane Mothe	IRIT, Université de Toulouse, France
Henning Müller	University of Applied Sciences Western Switzerland (HES-SO), Switzerland
Jian-Yun Nie	Université de Montréal, Canada
Eric SanJuan	University of Avignon, France
Giuseppe Santucci	Sapienza University of Rome, Italy
Jacques Savoy	University of Neuchâtel, Switzerland
Laure Soulier	Sorbonne University, France
Theodora Tsikrika	CERTH-ITI, Greece
Christa Womser-Hacker	University of Hildesheim, Germany

Past Members

Paul Clough	University of Sheffield, UK
Djoerd Hiemstra	Radboud University, The Netherlands
Jaana Kekäläinen	University of Tampere, Finland
Séamus Lawless	Trinity College Dublin, Ireland
Carol Peters (Steering Committee Chair 2000–2009)	ISTI-CNR, Italy
Emanuele Pianta	Centre for the Evaluation of Language and Communication Technologies (CELCT), Italy
Maarten de Rijke	University of Amsterdam, The Netherlands
Alan Smeaton	Dublin City University, Ireland

Supporters and Sponsors



Contents

Full Papers

Short Papers