Benchmarking Initiatives Testing the Algorithms of Natural Language and Multimedia Content Processing for Information Retrieval Purposes and Beyond

Maria Eskevich

#eScience2018
Amsterdam
29 Oct 2018

CLARIN

Definitions

- ***Benchmarking (Initiatives):*** comparison of different approaches to solve one task being defined in the same conditions using a set of performance metrics to define/rank the proposed solution.
- ***Natural Language Processing (NLP):*** analysis of content and information expressed in natural language data (text, audio) represented in machine processable form.
- ***Multimedia Content Processing:*** analysis of a the rich set of media combinations encompassing text, graphics, animation, sound, speech, image and video.
- ***Information Retrieval (IR):*** activity of finding the information across the resources/collections that is relevant to an information need expressed by a usere.
- ***Beyond …***

Benchmarking Initiatives types

- Initiative inspiration:
  - Bottom-um: community-driven
  - Top-down: use case scenarios provided by government/industry
- Approach to content selection:
  - Modality: Text/Audio/Video/Multimedia
  - Language group: e.g. languages of the region
  - Topic: e.g. biomedical content
  - Specific field: e.g. linked data, machine translation
- Type of tasks:
  - Benchmarking campaigns
  - Grand challenges
- Evaluation:
  - Predefined/annotated ground truth
  - Interactive evaluation and ground truth creation

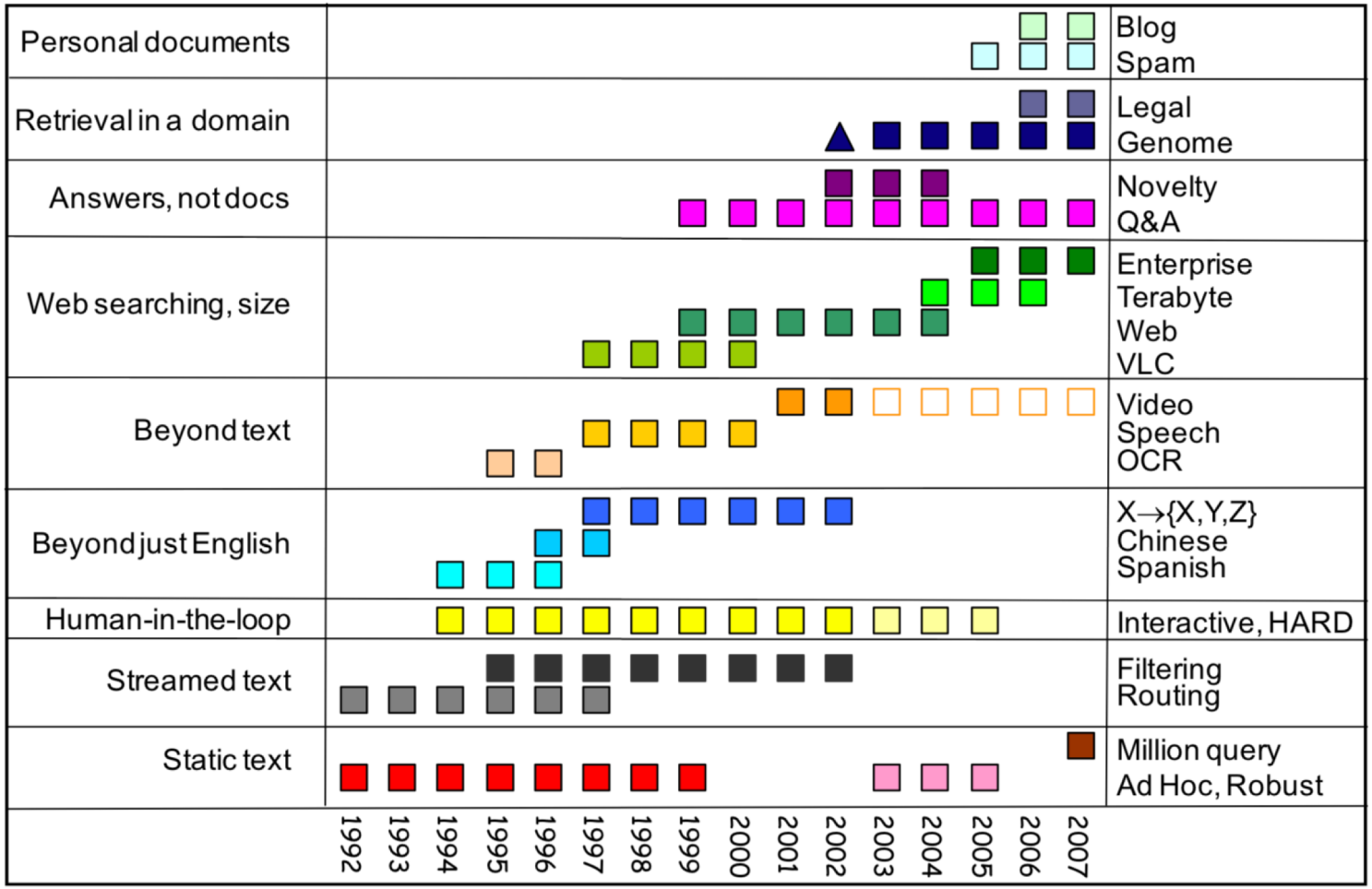
Modality: Text

Text Retrieval Conference (TREC)

- **Main orgasiser:** National Institute of Standards and Technology (NIST) and U.S. Department of Defense
- **Timeline**: started in 1992 as part of the TIPSTER Text program
- **Purpose**:
  - to support research in IR by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies
  - to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems
  - to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.
- **Focus**: large test collections
- **Audience**: industry, academia, and government
- **Webpage**: https://trec.nist.gov

Text Retrieval Conference (TREC)

Figure 2-2. TREC Tracks by Research Area, Title, and Year



From "Economic Impact Assessment of NIST's Text REtrieval Conference (TREC) Program" (2010)

TREC in 2018

- **CENTRE Track:** to develop and tune a reproducibility evaluation protocol for IR.
- **Common Core Track:** to investigate new methodologies for test collection construction (news documents).
- **Complex Answer Retrieval Track:** to develop systems that are capable of answering complex information needs by collating relevant information from an entire corpus.
- **Incident Streams Track:** to automatically process social media streams during emergency situations with the aim of categorizing information and aid requests made on social media for emergency service operators.
- **News Track:** to develop test collections that support the search needs of news readers and news writers in the current news environment.
- **Real-Time Summarization Track:** to construct real-time update summaries from social media streams in response to users' information needs.

Conference and Labs of the Evaluation Forum (CLEF)

- **Main orgasiser:** community-driven (Europe)
- **Timeline**: started in 2010
- **Purpose**:

- **Focus**: multimodality, multiliguality, vizualisation
- **Audience**: academia, industry
- **Webpage**: http://www.clef-initiative.eu, http://clef2018.clef-initiative.eu

CLEF in 2018

- **LifeCLEF:** to boosting biodiversity informatics related challenges (location-based species recommendation, bird species identification from bird calls and songs,experts vs. machines identification quality)
- **Lab on Digital Text Forensics:** to achieve cross-domain authorship attribution, to deanomyize authors, to identify an author's traits based on their writing style.
- **eHealth**: to support the development of techniques to aid laypeople, clinicians and policy-makers in easily retrieving and making sense of medical content to support their decision making.
- **Multilingual Cultural Mining and Retrieval:** to mine the social media sphere surrounding cultural events such as festivals.
- **Early Risk Prediction on the Internet (eRisk): to** detect early traces of depression/anorexia in the online user generated content.

CLEF in 2018

- **Automatic Identification and Verification of Political Claims (CheckThat!):** to foster the development of technology capable of both spotting and verifying check-worthy claims in political debates in English and Arabic.
- **Dynamic Search for Complex Tasks (DynSe):** to develop algorithms which interact dynamically with user (or other algorithms) towards solving a task, and evaluation methodologies to quantify their effectiveness.
- **Evaluation of Personalised Information Retrieval (PIR-CLEF):** to facilitate comparative evaluation of PIR.
- **Multimedia Retrieval in CLEF (ImageCLEF):** to advance lifelogs summarization and retrieval, bio-medical image search, medical question answering based on the visual image content.

Modality: Video

TREC Video Retrieval Evaluation (TRECVID)

- **Main orgasiser:** National Institute of Standards and Technology (NIST) and U.S. Department of Defense
- **Timeline**:
  - started in 2001 as a track in TREC
  - since 2003 independent initiative
- **Purpose**:
  - to develop and evaluate realistic system tasks and test collections
  - to use unfiltered data
  - to focus on relatively high-level functionality (e.g. interactive search)
  - to measure against human abilities
- **Focus**: content-based video analysis, retrieval, detection, etc
- **Audience**: industry, academia, and government
- **Webpage**: https://trecvid.nist.gov

TRECVID in 2018

- **Ad-hoc Video Search (AVS):** to model the end user search use-case, who is looking for segments of video containing persons, objects, activities, locations, etc. and combinations of the former.
- **Activities in Extended Video (ActEv):** to detect automatically activities for a multi-camera streaming video environment for both forensic applications and for real-time alerting.
- **Instance search (INS):** to find more video segments of a certain specific person, object, or place, given a visual example, in a given video collection.
- **Streaming Multimedia Knowledge Base Population (SMKBP):** to analyse each incoming information item (~100k items in total from a variety of genres, both formal (e.g.news) and informal (e.g., social media, blogs)) and to produce a set of structured representations (knowledge elements) about events, sub-events or actions, entities, relations, locations, time, and sentiments (beliefs).
- **Social-media video storytelling linking (LNK):** to create visual timelines/ summarization using collaborative videos, images and texts available from professional media and social-media users.
- **Video to Text Description (VTT):** to create automatic annotation of videos using natural language text descriptions.

Modality : Multimedia

MediaEval Benchmarking Initiative for Multimedia Evaluation (MediaEval)

- **Main orgasiser:** community-driven benchmark (Europe)
- **Timeline**:
  - Started as VideoCLEF track at CLEF
  - Since 2010 independent initiative
- **Purpose**:
  - to explore the social and human aspects of multimedia access and retrieval.
- **Focus**: The "multi" in multimedia: speech, audio, visual content, tags, users, context.
- **Audience**: academia, industry
- **Webpage**: http://www.multimediaeval.org

MediaEval in 2018

- **Emergency Response for Flooding Events:** to combine the information inherent in satellite images and social multimedia content in order to provide a more comprehensive view of disaster events.
- **Medico Multimedia Task:** to develop classifiers that minimize the necessary resources (processing time, training data) when processing medical multimedia data for disease prediction.
- **AcousticBrainz Genre Task: Content-based music genre recognition from multiple sources:** to create a system that can automatically assign genre labels to the tracks.
- **Emotional Impact of Movies Task:** to elaborate systems designed to predict the emotional impact of movies (valence and arousal scores; location of fear in movies).
- **Predicting Media Memorability Task:** to predict multimedia content memorability.

MediaEval in 2018

- **Human Behavior Analysis Task: No-Audio Multi-Modal Speech Detection in Crowded Social Settings:** to analyse automatically the conversational dynamics in large unstructured social gatherings such as networking or mingling events.
- **GameStory: Video Game Analytics Challenge:** to investigate ways to summarize how e-sport matches ramp up, evolve and play out over time.
- **Recommending Movies Using Content: Which content is key?** to create an automatic system that can predict the average ratings that users assign to movies and also the rating variance.
- **Pixel Privacy Task:** to create technology that invisibly changes or visibly enhances images in such a way that it is no longer possible to automatically infer the location at which they were taken.
- **NewsREEL Multimedia: News recommendation with image/text content:** to explore the relationship between images accompanying news articles, and the number of times these articles are clicked by users.

Language group

NII Testbeds and Community for Information access Research (NTCIR)

- **Main orgasiser:** Japan Society for Promotion of Science (JSPS), National Center for Science Information Systems (NACSIS), JSPS and Research Center for Information Resources at National Institute of Informatics (RCIR/NII)
- **Timeline**:
  - Started as 1998
- **Purpose**:
  - to investigate evaluation methods of Information Access techniques and methods for constructing a large-scale data set reusable for experiments.
  - to shift from document retrieval to "information" retrieval and technologies to utilizing information in the documents.
  - To investigation for realistic evaluation, including evaluation methods for summarization, multigrade relevance judgments and single-numbered averageable measures for such judgments.
- **Focus**: information retrieval with Japanese or other Asian languages and cross-lingual information retrieval.
- **Audience**: academia, industry
- **Webpage**: http://research.nii.ac.jp/ntcir/index-en.html

NTCIR in 2018

- **Lifelog:** to explore known-item search, exploratory search supported by lifelogs knowledge mining, to annotate multimodal data with human activities.
- **Open Live Test for Question Retrieval (OpenLiveQ-2**): to evaluate question answering given a query and a set of questions with their answers, while returning a ranked list of questions.
- **Question answering task for fact checking using Japanese regional assembly minutes (QALab-PoliInfo):** to extract structured data on the opinions of assemblymen, and the reasons and conditions for such opinions, from Japanese regional assembly minutes.
- **Short Text Conversation Task:** to estimate customer advance towards the Problem Solved state as distribution of overall subjective scores given a helpdesk-customer dialogue.
- **Ad hoc web search:** information retrieval for Chinese data
- **Fine-grained numeral understanding in financial social media data**

Forum for Information Retrieval Evaluation (FIRE)

- **Main orgasiser:** Governmental funding and community-driven
- **Timeline**:
  - Started as 2008
- **Purpose**:
  - to encourage research in Indian language Information Access technologies by providing reusable large-scale test collections for ILIR experiments
  - to provide a common evaluation infrastructure for comparing the performance of different IR system
  - to investigate evaluation methods for Information Access techniques and methods for constructing a reusable large-scale data set for ILIR experiments.
- **Focus**: multilingual information access.
- **Audience**: academia, industry
- **Webpage**: http://fire.irsi.res.in/fire

FIRE in 2018

- **Event Extraction from Newswires and Social Media Text in Indian Languages (EventXtract-IL):** to identify the event, event span, and further the cause and effects of a given event.
- **Information Extractor for Conversational Systems in Indian Languages (IECSIL):** Named Entity Recognition (NER) and Relation extraction
- **Information Retrieval from Microblogs during Disasters (IRMiDis):** to identify factual or fact-checkable tweets, and supporting news articles for fact-checkable tweets
- **Multilingual Author Profiling on SMS (MAPonSMS):** to determine author's gender and age (both individually and jointly) based on the multilingual input.
- **Verb Phrase Translation in English and Indian languages (VPT-IL):** to translate Verb Phrases from English to Tamil and Hindi to Tamil.
- **Indian Native Language Identification (INLI):** to identify the native language of the writer from the given Text/XML file which contains a set of Facebook comments in English language.

Specific Topic

International Workshop on Semantic Evaluation (SemEval)

- **Main orgasiser:** community-driven
- **Timeline**:
  - Started in 1998 as SenseEval (word sense disambiguation task)
  - Since 2007 independent initiative
- **Purpose**:
  - to evaluate semantic analysis systems
- **Focus**: what is necessary to compute in meaning as expressed in natural language
- **Audience**: academia, industry
- **Webpage**: http://alt.qcri.org/semeval2018/

SemEval in 2018

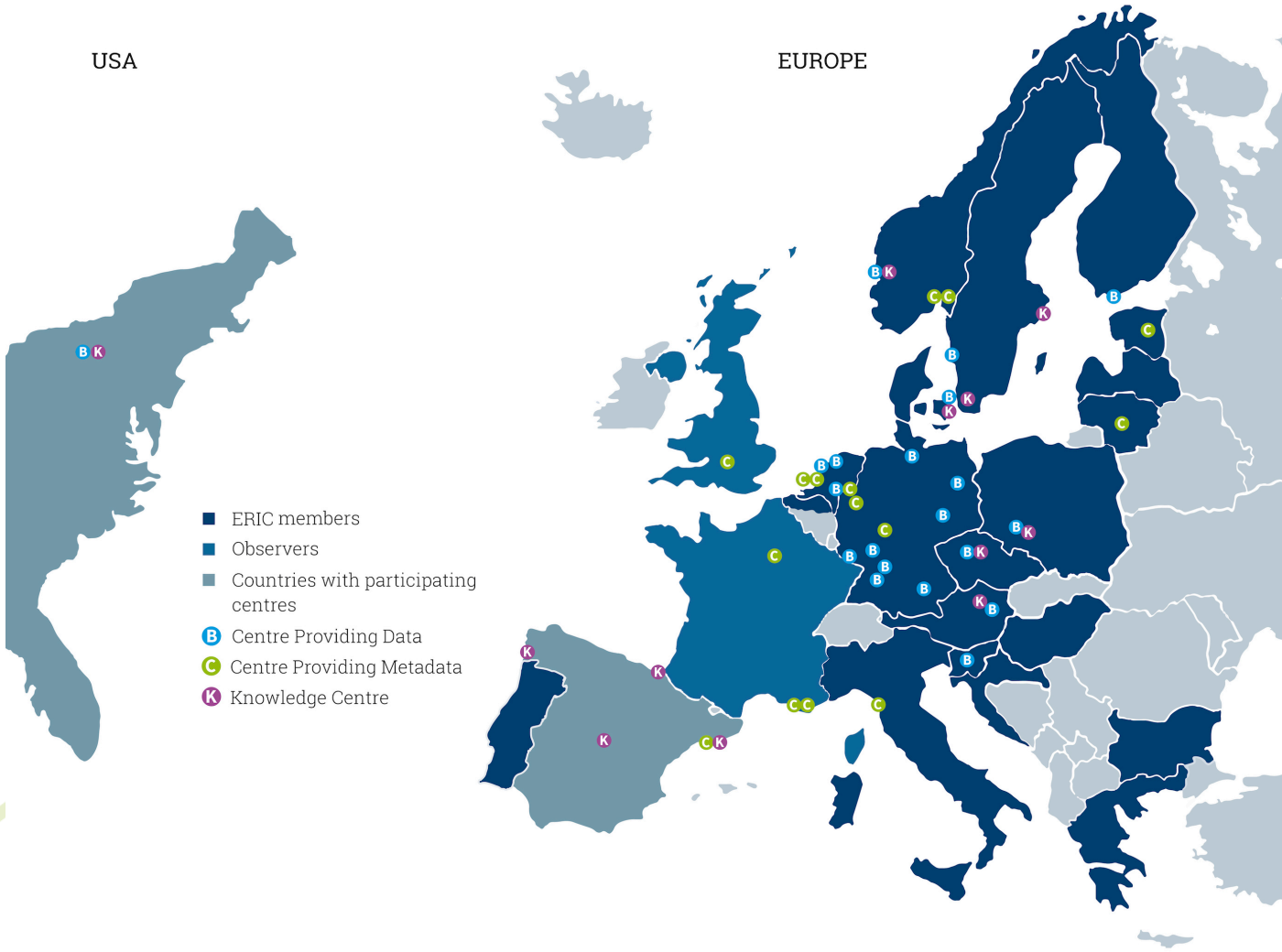
- **Affect and Creative Language in Tweets:** to automatically determine the intensity of emotions (E) and intensity of sentiment (aka valence V) of the tweeters from their tweets.
- **Coreference:** to solve coreference resolution task in the context of multiparty dialogues
- **Information Extraction:** to extract structured information from collections in specific domains such as medical reports, scientific papers, cybersecurity reports
- **Lexical Semantics:** to discover hypernym in general purpose and domain-specific settings
- **Reading Comprehension and Reasoning:** to advance automatic question answering and reasoning

Going beyond…

CLARIN ERIC in members and centres

A consortium of:

- 20 members: AT, BG, CZ, DE, DK, DLU, EE, FI, GR, HR, HU, IT, LT, LV, NL, NO, PL, PT, SE, SI
- 2 observers: FR, UK;
- >40 centres



CLARIN in seven bullets

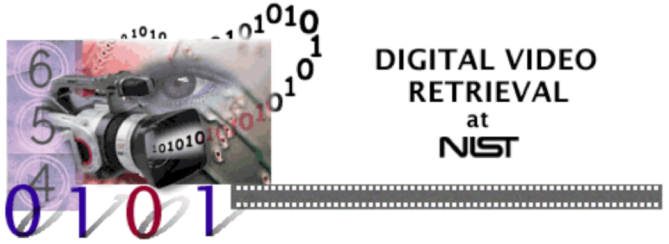
- **CLARIN** is the Common Language Resources and Technology Infrastructure
- **ESFRI** ERIC status since 2012, Landmark since 2016
- that provides easy and sustainable access for scholars in the **humanities and social sciences** and beyond
- to **digital language data** (in written, spoken, video or multimodal form)
- and **advanced tools** to discover, explore, exploit, annotate, analyse or combine them, wherever they are located
- through a **single sign-on** environment
- and that serves as an ecosystem for **knowledge sharing**.

The field is so much broader…

Text REtrieval Conference (TREC)

...to encourage research in information retrieval from large text collections.

NTCIR



MediaEval Benchmark

FIRE 2018

Forum for Information Retrieval Evaluation

AVIGNON CLEF 2018

Conference and Labs for the Evaluation Forum