OPHÉLIE FRAISIER

INFORMATION RETRIEVAL – EVALUATION CAMPAIGNS

IRIS team – IRIT

29-02-2016

Report IRIT/RR–2016–04–FR

This document was written by Ophélie Fraisier during the first year of her thesis.

Supervisors:

- Mohand Boughanem
- Romaric Besançon
- Guillaume Cabanac
- Yoann Pitarch

Contact information:

Email: ophelie.fraisier@irit.fr
Page: https://www.irit.fr/~Ophelie.Fraisier
IRIS team page: https://www.irit.fr/IRIS-site/

Contents

Introduction

The notion of experimental evaluation can be traced back to 1958, with the Cranfield experiments. Nowadays, several evaluation campaigns perpetuate the philosophy behind this experiments: evaluating different systems on a common ground.

[Robertson, 2008] presents some of the early experiments in the field, all to way to the well-known TREC evaluation campaign (an history of TREC can be found in [Voorhees, 2007]), and [Voorhees, 2002] explores in more detail the fundamental assumptions and appropriate uses of the Cranfield paradigm.

[Sanderson, 2010] gives a complete view of test collection based evaluation, from the creation of the datasets to alternative data sources for evaluation, including possible flaws such as the bias created with pooling[1] or assessment consistency.

[1] The aim of pooling is to locate an unbiased sample of the relevant documents in a large test collection.

The question of assessment is also discussed in [Alonso and Mizzaro, 2012], where the authors experiment with crowdsourcing for relevance assessment. Their results tend to support the fact that crowdsourcing is a cheap, quick, and reliable alternative for relevance assessment.

[Kelly, 2007] for its part presents the methods for evaluating retrieval systems with users whilst avoiding bias. This covers a wide range of questions, such as "how to recruit subjects?", "do the evaluation have to take part in a lab?", "in which order do the users have to test the system?", ...

Finally it is good to know that this document was written during a thesis on sentiment analysis on social networks, and is therefore focused on this particular field.

Overview

This document will focus on tasks dealing with sentiment analysis on social medias.

Summary of dates for 2016 editions

	Start of registrations	End of evaluation	Conferences
NTCIR	Feb. 27, 2015	Feb. 01, 2016	Jun. 07 – 10, 2016
SemEval	Jun. 18, 2015	Jan. 31, 2016	Jun. 16 – 17, 2016
CLEF	Oct. 30, 2015	May 4, 2016	Sep. 05 – 08, 2016
TREC	Feb. 01, 2016	Aug. 31, 2016	Nov. 15 – 18, 2016

Interesting tasks for 2016 editions

- TREC: Real-Time Summarization Track – Explore techniques for constructing real-time update summaries from social media streams in response to users' information needs.
- SemEval: Sentiment Analysis in Twitter, Aspect-Based Sentiment Analysis, Detecting Stance in Tweets and Determining Sentiment Intensity of English and Arabic Phrases tasks.
- CLEF: Cultural Microblog Contextualization – the workshop aims at developing processing methods and resources to mine the social media sphere surrounding cultural events such as festivals.

Past tasks

- Microblog Track from TREC 2011 – 2015 (dataset available for 2011)
- Sentiment Track from SemEval 2015 (multiple datasets available online)
- Sentiment Analysis in Twitter from SemEval 2013 – 2014 (multiple datasets available online)
- "Fouille d'opinion, de sentiment et d'émotion dans des messages postés sur Twitter" from DEFT 2015 (dataset available online)

Text REtrieval Conference (TREC)[2]

[2] Text REtrieval Conference (TREC), 2016a. URL http://trec.nist.gov; and Text Retrieval Conference on Wikipedia, 2016b. URL https://en.wikipedia.org/wiki/Text_Retrieval_Conference

[3] National Institute of Standards and Technology

[4] Official website: http://www.itl.nist.gov/iaui/894.02/related_projects/tipster/

Presentation

TREC is co-sponsored by the NIST[3] and U.S. Department of Defense and was started in 1992 as part of the TIPSTER Text program.[4]

It aims to improve communication and transfer of technology between academia, industries and governments by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.

For each TREC:

- NIST provides a test set of documents and questions.
- Participants run their own retrieval systems on the data, and return to NIST a list of the retrieved top-ranked documents.
- NIST pools the individual results, judges the retrieved documents for correctness, and evaluates the results.
- The TREC cycle ends with a workshop that is a forum for participants to share their experiences.

The TREC test collections and evaluation software are available to the retrieval research community at large, so organizations can evaluate their own retrieval systems at any time.

TREC has also sponsored the first large-scale evaluations of the retrieval of non-English (Spanish and Chinese) documents, retrieval of recordings of speech, and retrieval across multiple languages. TREC has also introduced evaluations for open-domain question answering and content-based retrieval of digital video.

Tracks for TREC 2016:

- Clinical Decision Support Track: Investigate techniques for linking medical cases to information relevant for patient care
- Contextual Suggestion Track: Investigate search techniques for complex information needs that are highly dependent on context and user interests.
- Dynamic Domain Track: Investigate domain-specific search algorithms that adapt to the dynamic information needs of professional users as they explore in complex domains.
- LiveQA Track: Generate answers to real questions originating from real users via a live question stream, in real time.
- OpenSearch Track *(first year)*: Explore an evaluation paradigm for IR that involves real users of operational search engines. For this first year of the track the task will be ad hoc Academic Search.

- **Real-Time Summarization Track: Explore techniques for constructing real-time update summaries from social media streams in response to users' information needs.**
- Tasks Track: Test whether systems can induce the possible tasks users might be trying to accomplish given a query.
- Total Recall Track: Evaluate methods to achieve very high recall, including methods that include a human assessor in the loop.

Tasks for the Real-time summarization track

The tasks for this track are not finalized yet, or not publicly available.

Schedule for TREC 2016

Feb. 01, 2016	Submission of applications to participate in TREC 2016
Feb. 25, 2016	Password communicated to the participants
Mar. 1, 2016	Distribution of document disks used in some existing TREC collections to the participants
Jul. – Aug., 2016	Results submission deadline (some tracks may have a late spring submission deadline)
Sep. 30, 2016	Relevance judgments and individual evaluation scores due back to participants
Nov. 15–18, 2016	TREC 2016 conference at NIST in Gaithersburg, Md. USA.

Past editions

From 2011 to 2015: the **Microblog Track** examined the nature of real-time information needs and their satisfaction in the context of microblogging environments such as Twitter.

Dataset from 2011: Tweets2011, approximately 16 million tweets sampled between January 23rd and February 8th, 2011. The corpus is designed to be a reusable, representative sample of the twittersphere - i.e. both important and spam tweets are included.

The Tweets2011 corpus is unusual in that what you get is a list of tweet identifiers, and the actual tweets are downloaded directly from Twitter, using the open-source twitter-tools. However, to obtain the lists of tweets to be downloaded (i.e. the "tweet lists"), a data usage agreement must be signed. Once signed, the agreement must be emailed back to NIST, who will provide you with a username/password to download the tweet lists (in the form of a .tar.gz file).

Semantic Evaluation (SemEval)[5]

[5] International Workshop on Semantic Evaluation (SemEval-2016), 2016. URL http://alt.qcri.org/semeval2016; and SemEval on Wikipedia, 2016. URL https://en.wikipedia.org/wiki/SemEval

Presentation

SemEval is an ongoing series of evaluations of computational semantic analysis systems whose evaluation workshop is held during the yearly SEM conference[6] or the NAACL conference.[7]

It evolved from the Senseval word sense evaluation series (1998, 2001, 2004), were focused on word sense disambiguation. Beginning with the fourth workshop, SemEval-2007, the nature of the tasks evolved to include semantic analysis tasks outside of word sense disambiguation.

[6] Official website: https://sem.org/CONF-AC-TOP.asp

[7] Official website: http://naacl.org/naacl-hlt-2016/

Tracks (2016)

- Textual Similarity and Question Answering (Tasks 1 to 3)
- **Sentiment Analysis (Tasks 4 to 7)**
- Semantic Parsing (Tasks 8 and 9)
- Semantic Analysis (Tasks 10 to 12)
- Semantic Taxonomy (Tasks 13 and 14)

Tasks for the Sentiment Analysis track

Task 4: Sentiment Analysis in Twitter

Subtasks:

- (rerun) Message Polarity Classification: Given a tweet, predict whether the tweet is of positive, negative, or neutral sentiment.
- (partially new) Tweet classification according to a two-point scale: Given a tweet known to be about a given topic, classify whether the tweet conveys a positive or a negative sentiment towards the topic.
- Tweet classification according to a five-point scale: Given a tweet known to be about a given topic, estimate the sentiment conveyed by the tweet towards the topic on a five-point scale.
- Tweet quantification according to a two-point scale: Given a set of tweets known to be about a given topic, estimate the distribution of the tweets across the Positive and Negative classes.

- Tweet quantification according to a five-point scale: Given a set of tweets known to be about a given topic, estimate the distribution of the tweets across the five classes of a five-point scale.

Task 5: Aspect-Based Sentiment Analysis

Subtasks:

- Sentence-level ABSA: Given an opinionated document about a target entity, identify all the opinion tuples with the following types of information:
  - Aspect Category Detection (identify every entity E and attribute A pair towards which an opinion is expressed in the given text)
  - Opinion Target Expression (OTE) (extract the linguistic expression used in the given text to refer to the reviewed entity E of each E#A pair)
  - Sentiment Polarity (each identified E#A, OTE tuple has to be assigned a polarity labels)
- Text-level ABSA: Given a set of customer reviews about a target entity, identify a set of tuples that summarize the opinions expressed in each review.
- Out-of-domain ABSA: test of the systems in a previously unseen domain for which no training data will be made available (supported for the French language).

Datasets: [8] English (2 domains, fine-grained human annotations), Arabic, Chinese (2 domains), Dutch (2 domains), French, Russian (no information on annotation), Spanish (no information on annotation), Turkish (2 domains, no information on annotation).

[8] One domain-specific dataset manually annotated per language, unless otherwise specified.

Task 6: Detecting Stance in Tweets

Subtasks:

- Supervised framework: stance[9] towards five targets: "Atheism", "Climate Change is a Real Concern", "Feminist Movement", "Hillary Clinton", and "Legalization of Abortion". Dataset: 2900 labeled training data instances for the five targets.
- Weakly supervised framework: stance towards one target "Donald Trump". No training data, large set of tweets associated with "Donald Trump" but it is not labeled for stance.

[9] Classes: FAVOR, AGAINST, NONE

Task 7: Determining Sentiment Intensity of English and Arabic Phrases

Objective: to test an automatic system's ability to predict a sentiment intensity score for a word or a phrase.[10]

[10] Including categories known to be challenging for sentiment analysis (negators, modals, intensifiers and diminishers)

Datasets:

- General English Sentiment Modifiers Set: phrases formed by combining a word and a modifier (negator, auxilary verb, degree adverb or combination of those) and single word terms (which are part of the multi-word phrases, as separate entries).
- English Twitter Mixed Polarity Set: phrases made up of opposite polarity terms and single word terms (which are part of the multi-word phrases[11], as separate entries). This allows the evaluation to determine how good the automatic systems are at determining

[11] Drawn from a corpus of tweets, may include a small number of hashtag words and creatively spelled words.

sentiment association of individual words as well as how good they are at determining sentiment of phrases formed by their combinations.

- Arabic Twitter Set: single words and phrases combining a negator and a word commonly found in Arabic tweets.

Schedule for SemEval 2016

Jun. 30, 2015	Trial data ready
Aug. 30, 2015	Training data ready
Jan. 10 – Jan. 31, 2016	Evaluation
Feb. 26, 2016	Paper submission due
Mar. 16, 2016	Paper reviews due
Mar. 18, 2016	Author Notifications
Mar. 25, 2016	(For shepherded papers only) Revised submission due
Apr. 1, 2016	(For shepherded papers only) Shepherded author notification
Apr. 7, 2016	Camera ready due
Jun. 16-17, 2016	SemEval workshop at NAACL

Past editions

- Sentiment Track (SemEval 2015, multiple datasets available online)
- Sentiment Analysis in Twitter (SemEval 2013 – 2014, multiple datasets available online)

Conference and Labs of the Evaluation Forum (CLEF)[12]

[12] The CLEF Initiative, 2016. URL http://www.clef-initiative.eu

Presentation

The CLEF Initiative[13] is a self-organized body which promotes research and development by providing an infrastructure for:

[13] Formerly known as Cross-Language Evaluation Forum.

- multilingual and multimodal system testing, tuning and evaluation;
- investigation of the use of unstructured, semi-structured, highly-structured, and semantically enriched data in information access;
- creation of reusable test collections for benchmarking;
- exploration of new evaluation methodologies and innovative ways of using experimental data;
- discussion of results, comparison of approaches, exchange of ideas, and transfer of knowledge.

The CLEF Initiative is structured in two main parts:

1. a series of Evaluation Labs, i.e. laboratories to conduct evaluation of information access systems and workshops to discuss and pilot innovative evaluation activities;
2. a peer-reviewed Conference on a broad range of issues, including
   - investigation continuing the activities of the Evaluation Labs;
   - experiments using multilingual and multimodal data;
   - research in evaluation methodologies and challenges.

CLEF 2016 will be hosted by the Computer Science Department of the School of Sciences and Technology of the University of Évora, Portugal, 5-8 September 2016.

Labs for 2016:

- CLEF eHealth: the goals are to develop processing methods and resources in a multilingual setting to enrich difficult-to-understand eHealth textx, and provide valuable documentation.
- ImageCLEF: the task tackles different aspects of the annotation problem: Image Annotation, ImageCLEFmed: The Medical task, Handwritten Document Retrieval.
- LifeCLEF: it aims at evaluating multimedia analysis and retrieval techniques on biodiversity data for species identification.

- LL4IR (Living Lab for IR): the main goal is to provide a benchmarking platform for researchers to evaluate their ranking systems in a live setting with real users in their natural task environment.
- NEWSREEL (News Recommandation Evaluation Lab): the lab will address the challenge of real-time news recommendation.
- PAN (uncovering Plagiarism, Authorship and Social Softare Misuse): the main goal to provide for sustainable and reproducible evaluations, to get a clear view of the capabilities of state-of-the-art-algorithms.
- SBS (Social Book Search): the goal is to investigate techniques to support users in complex book search tasks that involve more than just a query and results list.
- **CMC (Cultural Microblog Contextualization): the workshop aims at developing processing methods and resources to mine the social media sphere surrounding cultural events such as festivals.**

Tasks for the CMC lab 2016

Task 1: Cultural Multilingual microblog contextualization based on WikiPedia

Objective: Given a microblog announcing some cultural event, provide in real time a summary[14] extracted from the wikipedia and readable on a small device that provides extensive background about this event.

[14] Not exceeding 500 words, composed of passages from a provided Wikipedia corpus.

Dataset: The document collection will be rebuilt based on various Wikipedias from next November 2015.

Task 2: Cultural MicroBlog Search based on WikiPedia entities

Objective: Given a cultural entity as a set of WikiPedia pages[15] provide a double extensive summary of relevant microblogs from insiders and outsiders.

[15] Set of places to visit, artists to see on stage, festivals of interest, ...

Subtasks:

- Task 2a: Retrieval of relevant microblogs for an entity (described by its wikipedia page)
- Task 2b: Summarization of the most informative tweets (and comparison to manually built summaries)

Datasets: datasets created within the GAFES project, pool of 10 million microblogs with their meta-information, images and two steps crawl of urls inside microblogs, as well as ground truth for the evaluation + clean simplified xml dump of wikipedia easy to index and to process with state of the art NLP tools is made available to participants.

Task 3: TimeLine illustration based on Microblogs

Objective: link the event programme elements of a given programme to related microblogs.

Dataset: Microblogs will be provided with their timestamps.

Schedule

These are the dates provided for the "Labs" part of CLEF 2016.

Sep. 1, 2015	Final lab proposals
Sep. 14, 2015	Notification of lab acceptance
Oct. 30, 2015	Labs registration opens
Apr. 22, 2016	Registration closes
May 4, 2016	End Evaluation Cycle
May 25, 2016	Submission of Participant Papers [CEUR-WS]
Jun. 3, 2016	Submission of Lab Overviews [LNCS]
Jun. 10, 2016	Notification of Acceptance Lab Overviews [LNCS]
Jun. 17, 2016	Camera Ready Copy of Lab Overviews [LNCS] due
Jun. 17, 2016	Notification of Acceptance Participant Papers [CEUR-WS]
Jul. 1, 2016	Camera Ready Copy of Participant Papers and Extended Lab Overviews [CEUR-WS] due
Sep. 5 – 8, 2016	CLEF 2016 at the University of Évora, Portugal

Forum For Information Retrieval Evaluation (FIRE)[16]

[16] Forum for Information Retrieval Evaluation (FIRE), 2016. URL http://fire.irsi.res.in

Presentation

FIRE started in 2008 with the aim of building a South Asian counterpart for TREC, CLEF and NTCIR. It has expanded to include new domains like plagiarism detection, legal information access, mixed script information retrieval and spoken document retrieval to name a few.

FIRE 2015 introduced a peer-reviewed conference track and was held in Dhirubhai Ambani Institute of Information and Communication Technology, Gandhinagar, India.

No information is available concerning FIRE 2016.

The topics including in FIRE 2015 were:

- Information Retrieval
- Multilingual and cross-lingual Information Access
- Domain specific Information Access
- Natural Language Processing
- Interactive Information Retrieval / Human Computer Interaction
- Computational Linguistics
- Semantic Web
- Other related fields (Social media analysis, Digital Library, Enterprise search, etc)

No information seems available for FIRE 2016.

NII Testbeds and Community for Information Research (NTCIR)[17]

[17] NII Testbeds and Community for Information Research (NTCIR), 2016. URL http://research.nii.ac.jp/ntcir/index-en.html

Présentation

Since 1997, the NTCIR project has promoted research efforts for enhancing Information Access (IA) technologies such as Information Retrieval (IR), Text Summarization, Information Extraction (IE), and Question Answering (QA) techniques.

Its general purposes are to:

- Offer research infrastructure that allows researchers to conduct a large-scale evaluation of IA technologies
- Form a research community in which findings based on comparable experimental results are shared and exchanged
- Develop evaluation methodologies and performance measures of IA technologies.

Tasks for 2016:

- Search Intent and Task Mining: aims to explore and evaluate the technologies of understanding user intents behind the query and satisfying different user intents.
- Medical Natural Language Processing for Clinical Document: participants are supposed to assigning a suitable diagnosis and the corresponding disease code to a clinical case in Japanese.
- Mobile Information Access: participants are required to generate a two-layered summary in response to a given query that fits screens of mobile devices instead of ten blue links.
- Spoken Query and Spoken Document Retrieval: evaluates spoken document retrieval from spontaneously spoken query.
- Temporal Information Access: aims to foster research in temporal information access with the following subtasks: Temporal Intent Disambiguation and Temporally Diversified Retrieval.
- Mathematical Information Retrieval: aims to develop a test collection for evaluating retrieval using queries comprised of keywords and formulae, in order to facilitate and encourage research in mathematical information retrieval (MIR) and its related fields.
- Lifelog Task: aims to begin the comparative evaluation of information access and retrieval systems operating over personal lifelog data, with 2 subtasks:

  - Lifelog Semantic Access Task: known-item search task that can be undertaken in an interactive or automatic manner.
  - Lifelog Insight Task: develop tools and interfaces that facilitate filtering and provide for efficient/effective means of visualisation of the data.

- QA Lab for Entrance Exam: The goal is to investigate the real-world complex Question Answering (QA) technologies using Japanese university entrance exams and their English translation on the subject of "World History".
- Short Text Conversation Task ("STC"): improve natural language conversation between human and computer, with a simplified version of the problem: one round of conversation formed by two short texts, with the former being an initial post from a user and the latter being a comment given by the computer.

Schedule for NTCIR 12 (2016)

Jul. 31 2015	Task Registration Due
Jul. 01 2015	Document Set Release
Jul. – Dec. 2015	Dry Run
Sep. 2015 – Feb. 2016	Formal Run
Feb. 01 2016	Evaluation Results Return
Feb. 01 2016	Early draft Task Overview Release
Mar. 01 2016	Draft participant paper submission Due
May 01 2016	All camera-ready paper for the Proceedings Due
Jun. 07–10 2016	NTCIR-12 Conference in NII, Tokyo, Japan

DÉfi Fouille de Texte (DEFT)[18]

[18] DÉfi Fouille de Texte, 2016. URL https://deft.limsi.fr

Presentation

DEFT was created in 2005 to provide a French-speaking evaluation campaign in text mining.

Text mining aims to automatically extract and organize information from text. 2 types of methods are used to reach this goal:

- those based on experts' knowledge
- those based on automatic supervised learning

No information was found about DEFT 2016.

Past editions

DEFT 2015 could be useful for our research: "fouille d'opinion, de sentiment et d'émotion dans des messages postés sur Twitter"

Datasets available here: https://deft.limsi.fr/2015/corpus.fr.php?lang=fr

Articles available here: http://www.atala.org/taln_archives/ateliers/2015/DEFT/

Bibliography

The CLEF Initiative, 2016. URL http://www.clef-initiative.eu.

DÉfi Fouille de Texte, 2016. URL https://deft.limsi.fr.

Forum for Information Retrieval Evaluation (FIRE), 2016. URL http://fire.irsi.res.in.

International Workshop on Semantic Evaluation (SemEval-2016), 2016. URL http://alt.qcri.org/semeval2016.

NII Testbeds and Community for Information Research (NTCIR), 2016. URL http://research.nii.ac.jp/ntcir/index-en.html.

SemEval on Wikipedia, 2016. URL https://en.wikipedia.org/wiki/SemEval.

Text REtrieval Conference (TREC), 2016a. URL http://trec.nist.gov.

Text Retrieval Conference on Wikipedia, 2016b. URL https://en.wikipedia.org/wiki/Text_Retrieval_Conference.

Omar Alonso and Stefano Mizzaro. Using crowdsourcing for TREC relevance assessment. *Inf. Process. Manag.*, 48(6):1053–1066, 2012. DOI: 10.1016/j.ipm.2012.01.004.

Diane Kelly. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Found. Trends® Inf. Retr.*, 3(1—2):1–224, 2007. DOI: 10.1561/1500000012.

S. Robertson. On the history of evaluation in IR. *J. Inf. Sci.*, 34(4):439–456, 2008. DOI: 10.1177/0165551507086989.

Mark Sanderson. Test Collection Based Evaluation of Information Retrieval Systems. *Found. Trends® Inf. Retr.*, 4(4):247–375, 2010. DOI: 10.1561/1500000009.

Ellen M. Voorhees. The Philosophy of Information Retrieval Evaluation. In *Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems*, CLEF '01, pages 355–370. Springer-Verlag, 2002. ISBN 3-540-44042-9. URL http://dl.acm.org/citation.cfm?id=648264.753539.

Ellen M. Voorhees. TREC: Continuing information retrieval's tradition of experimentation. *Commun. ACM*, 50(11):51, 2007. DOI: 10.1145/1297797.1297822.