Computational fact-checking: Problems, state of the art, and perspectives

Julien Leblay, Ioana Manolescu, Xavier Tannier

Computational fact-checking: problems, state of the art, and perspectives

Tutorial proposal for WWW (the Web Conference) 2018

Julien Leblay, Ioana Manolescu, Xavier Tannier

Keywords: fact-checking, reasoning, data management, natural language processing, claim extraction

Organizers

Julien Leblay, Research Scientist at the National Institute of Advanced Industrial Science and Technology (AIST) in Tokyo, Japan. https://staff.aist.go.jp/julien.leblay/

- As part of the Artificial Intelligence Research Center, his research interests cover data management and query processing in general, with a particular focus on applications to Web data, i.e., data typical found on web services and Open Data. From October 2013 to February 2015, Julien was a postdoctoral research assistant at the University of Oxford under the supervision of Michael Benedikt, where his work covered query optimization under constraints and access restrictions. He obtained his Ph.D. in Computer Science from the Université Paris-Sud and Inria, France, under the supervision of François Goasdoué and Ioana Manolescu. Prior to his academic career, he worked in industry for several years as a software engineer and consultant on topics ranging from machine translation to data integration.

Ioana Manolescu, Inria and Ecole Polytechnique, France, ioana.manolescu@inria.fr, http://pages.saclay.inria.fr/ioana.manolescu/

- Ioana Manolescu is a senior researcher at Inria Saclay and Ecole Polytechnique. She is the lead of the CEDAR INRIA team focusing on rich data analytics at cloud scale. She is a member of the PVLDB Endowment Board of Trustees, of the ACM SIGMOD Jim Gray PhD dissertation committee, and an associate editor for PVLDB. Recently, she has been the program chair of the Scientific and Statistical Data Management Conference 2016, and she is a co-chair of the upcoming IEEE ICDE conference in 2018. She has co-authored more than 130 articles in international journals and conferences, and contributed recently to a book on "Web Data Management" by S. Abiteboul, I. Manolescu, P. Rigaux, M.-C. Rousset and P. Senellart. Her main research interests include data models and algorithms for fact-checking, algebraic and storage optimizations for semistructured data and in particular data models for the Semantic Web, novel data models and languages for complex data management, and distributed architectures for complex large data.

Xavier Tannier, University Pierre and Marie Curie (UPMC, Paris 06), LIMICS -- xavier.tannier@upmc.fr http://xavier.tannier.free.fr/ (temporary)

- Xavier Tannier is a professor at University Pierre and Marie Curie and researcher at LIMICS since 2017. He was associate professor at University Paris-Sud and researcher at LIMSI-CNRS from 2007 to 2017. His main field of research lies in natural language processing and text mining in large collections of documents.

Abstract

The tremendous value of Big Data has been noticed of late also by the media, and the term "data journalism'' has been coined to refer to journalistic work inspired by digital data sources. A particularly popular and active area of data journalism is concerned with fact-checking. The term was born in the journalist community and referred the process of verifying and ensuring the accuracy of published media content; since 2012, however, it has increasingly focused on the

analysis of politics, economy, science, and news content shared in any form, but first and foremost on the Web (social and otherwise).
These trends have been noticed by computer scientists working in the industry and academia. Thus, a very lively area of digital content management research has taken up these problems and works to propose foundations (models), algorithms, and implement them through concrete tools. To cite just one example, Google has recognized the usefulness and importance of fact-checking efforts, by making an effort to index and show them next to links returned to the users (https://developers.google.com/search/docs/data-types/factcheck).

Our proposed tutorial:

1. Outlines the current state of affairs in the area of digital (or computational) fact-checking in newsrooms, by journalists, NGO workers, scientists and IT companies;
2. Shows which areas of digital content management research, in particular those relying on the Web, can be leveraged to help fact-checking, and gives a comprehensive survey of efforts in this area;
3. Highlights ongoing trends, unsolved problems, and areas where we envision future scientific and practical advances.

Topic and Relevance

Scope and depth of the tutorial

The tremendous value of Big Data has been noticed of late also by the media, and the term "data journalism'' has been coined to refer to journalistic work inspired by data sources. While data of some form is a natural ingredient of all reporting, the increasing volumes and complexity of digital data lead to a qualitative jump, where technical skills for working with data are stringently needed in journalistic work.

A particularly popular and active area of data journalism is concerned with **fact-checking**. The term was born in the journalist community; it referred to the task of identifying and checking factual claims present in media content, which dedicated newsroom personnel would then check for factual accuracy. The goal of such checking was to avoid misinformation, to protect the journal reputation and avoid legal actions.
Starting around 2012, first in the United States (http://factcheck.org), then in Europe, and soon after in all areas of the world, journalists have started to take advantage of modern technologies for processing content, such as text, video, structured and unstructured data, in order to automate, at least partially, the knowledge finding, reasoning, and analysis tasks which had been previously performed completely by humans. Since at about the same time the US and several European countries held major elections, at about the same time, the focus of fact-checking work shifted from verifying claims made by media outlets, toward the claims made by politicians and other public figures.
This trend also created a virtuous cycle with the parallel (but distinct) evolution toward asking Government Open Data, that is: the idea that governing bodies should share with the public precise information describing their functioning, so that the people have a means to assess the quality of their elected representation. Government Open Data became quickly available, in large volumes, e.g. through http://data.gov in the US, http://data.gov.uk in the UK, http://data.gouv.fr in France etc.; journalists turned out to be the missing link between the newly available data and comprehension by the public. Data journalism thus found one of its foremost and most useful incarnations in fact-checking based on digital content and tools; there are natural connections with investigative journalism, which also needs to identify, analyze and exploit complex databases. This has been illustrated most visibly in recent years by the Panama Papers (https://panamapapers.icij.org/) and Paradise Papers (https://www.icij.org/investigations/paradise-papers/) studies of tax evasion across the world.

Beyond journalists, concerned citizens, NGOs such as FactCheck.org, and scientists such as those running http://climatefeedback.org also joined the discussion; this has enlarged the scope of journalistic fact-checking, beyond politics, to issues related to health (medical scandals), the environment (pollution through dangerous pesticides, or the controversy over climate change, studied in particular by ClimateFeedback mentioned above) and many others.
Another parallel development is the massive production of **fake news** or massive manipulation through false news content. While (typically false) propaganda information is not novel, the Web and the social media, ampliflied by the so-called "echo chamber" and "filter bubble" effects, have taken its scale to a higher order of magnitude; fake news production is quasi industrial (see e.g., http://money.cnn.com/interactive/media/the-macedonia-story/?sr=fbCNN091317undefined0501PMStory)

These aspects being noticed by computer scientists which are also citizens and eager to contribute to the way modern society works, a very lively area of digital content management research has taken up these problems and works to propose foundations (models), algorithms, and implement them through concrete tools. The efforts have been many but scattered. Google, in particular, has recognized the usefulness and importance of fact-checking efforts, by making an effort to index and show them next to links returned by the users (https://developers.google.com/search/docs/data-types/factcheck).

If a fully automatic approach to fact-checking is beyond reach (and probably not even desirable), old and new fields of research from different domains are very relevant to help journalists deal with the change of scale in their missions.
We will survey recent efforts in the following areas of computer science, with a particular focus on those applied to, or pertinent for, the Web:

- **Data management**, in the sense of persisting data and querying it: journalists need it both for the claims which are made (typically publicly through the Web) and for the reference data sources which they can use in their verification (such as reference statistic datasets published by government agencies). Yet, our interactions with journalists and fact-checkers highlight that establishing repositories of persistent data is not an obvious thing for them, especially that they may want to store data files, but also links, establish interconnections, annotate the data etc. We will briefly review the kind of data sources they have to deal with, and existing techniques which data management (and in particular Web data management) may have to offer.
- **Data integration**, which allows exploiting together datasets of different origins and often independently produced, is also necessary, given the many such opportunities they encounter. This was in particular the case for the Panama and Paradise paper analyses (https://neo4j.com/blog/icij-neo4j-unravel-panama-papers/). We will review the data integration architectures [DS13] (mostly focusing on data warehouses, mediators, data spaces [FHM05] and data lakes [HKN+16]) and comment on their applicability to fact-checking scenarios we encountered. Still in a data integration scenario, a very relevant task is the selection of the best information sources to answer a specific query (in the classic scenarios) [CGY16,RDD16], or to check a specific claim (in a modern fact-checking scenario) [HZA+17]. In a related vein, **truth discovery** attempts to quantity the veracity of data when collected and merged from many, possible disagreeing, sources [DBS09a, DBS09b]. We will cover these techniques.
- **Text analysis and information extraction**, in particular through automated classification and learning, is gaining momentum as a way to cope with the huge number of documents published in social or mainstream media. In particular, in the context of the web, these techniques allow to go from unstructured or poorly structured text, to structured knowledge bases which lend themselves more easily to fact-checking answers. Text analysis can also be used to detect trends in news, to extract the source of claims [KTH+12,THM15,TV16] or to recognize rumors [CMP11]. There has also been some attempts to create end-to-end fact

validation systems collecting and monitoring facts, from online data, and looking for evidences to input claims [LGM+12, HSW+14]. News analysis has established itself as a research topic on its own, covering news clustering over time, causality detection between news events or credibility analysis.

- Many other fields of **natural language processing** are closely related: textual entailment is the task of comparing two portions of text and deciding whether the information contained in the first one can be implied from the second [DGM05,LSD13]. Stance detection aims at determining from a text whether it is in favor of a given target or against it [ARV+16]. Entity linking consists in connecting an entity mention that has been identified in a text to one of the known entities in a knowledge base [LSW15,SWH15].
- **Data mining and graph analysis** Data mining methods applied to structured and regular data enable powerful fine-granularity analysis of media claims; related work focus on very specific types of queries (e.g. checking that criminality rate has decreased during the mandate of M. X's as a mayor [HSW+14,WAL+14]) or on tracking exceptional events [BCL+17]. The context in which an information item is produced may hold valuable hints toward the trustworthiness of that information. Existing research on social network analytics may help identify communities of fake news producers, identify rumor spreaders etc.
- **Machine learning** is frequently leveraged to help classify published content according to their theme (topic), according to their likely trustworthiness or at least their "checkworthiness" [CMP11, GMG+16, HZA+17]. Journalists in particular strongly appreciate automated help to narrow the documents on which they should focus their verification effort. Fake news detection is now a very active field mobilizing a growing numbers of researchers, and is now the focus of international challenges (http://www.fakenewschallenge.org/, https://herox.com/factcheck etc.)
- **Image and video processing and classification**: pictures and videos are a very common way to disseminate fake news, either by lying about their provenance or date or creation, or by using photomontages. Tools have been created recently to help verifying the authenticity of a photo or a video (see for example the European project InVID).
- **Temporal and dynamic aspects of the above:** the news arena is by definition one of perpetual movement, and many areas of reality follow the same pattern; time is a natural dimension of all human activity. Facts can be true during a period of time and then become false. Also, many hoaxes are spread periodically, and many "news" can be false just because the fact they relate happened years ago.

Importance and timeliness of the tutorial

We believe the audience is likely to get many ideas for applications and research. It is timely in that the amount of ideas and research works involved is currently quite significant, and non-computer scientists (journalists mostly) are eager to be involved in designing and using novel tools for their work. We believe a missing piece of the puzzle to make it happen, is a dissemination effort on the needs on one hand and the available and future scientific tools on the other hand. This tutorial is an attempt to provide such a dissemination effort.

Relevance to WWW

As previously stated, most of the world's information is currently being exchanged through the Web. Automated or semi-automated tools to help fact-checking will necessarily draw from numerous areas of Web research, as outlined above. The conference has also opened in 2018 a satellite track on journalism, misinformation and fact-checking (https://www2018.thewebconf.org/call-for-papers/misinformation-cfp/) to which we hope to submit a companion paper for this tutorial.

Presenters' qualifications

J. Leblay and I. Manolescu have been among the first researchers considering fact-checking Web content from a data and knowledge management perspective, publishing a demonstration paper called "Fact-checking and analzing the Web" in the ACM SIGMOD conference in 2013 [GKK+13]. The demonstration was subsequently shown at the **Computation+Journalism** conference held in New York, in October 2014.
Subsequently, X. Tannier and I. Manolescu started in 2016 a national French research project called **ContentCheck** (https://team.inria.fr/cedar/contentcheck/) dedicated to content management models, algorithms and tools for journalistic fact-checking, in collaboration with **Les Décodeurs** (http://www.lemonde.fr/les-decodeurs/), a data journalism and fact-checking team of 12, part of Le Monde, France's leading national newspaper. During this project, I. Manolescu and X. Tannier have co-authored several publications, including a demonstration of a data integration prototype for data journalism at PVLDB 2016 [BCC+2016] and a tool for producing RDF Linked Open Data out of Excel and HTML tables, as an effort to improve reference sources to be used for fact-checking [CMT17]. Meanwhile, J. Leblay has continued work on foundations and models for explaining, analyzing and checking claims in context, in particular authoring an article on this topic in the AAAI conference 2017 [Leblay17] as well as a recent demonstration at ACM CIKM [LCL17].
X. Tannier has many years of experience on conducting text analysis and NLP projects together with major media actors, notably AFP (Agence France Presse). More generally, he has also worked extensively on event extraction from journalistic content, text classification etc.
The authors are part of a bilateral collaboration team co-funded by the JSPS (Japan Society for the Promotion of Science) and Inria, aimed at developing models and tools to find explanations for claims using Web data.

Going beyond our scientific activity, over the last few years, we have met with many enthusiasts in the area (journalists, bloggers, computer scientists, hackers, or some mix of the above), enabling us to learn about other people's interesting problems and efforts in this area, which is extremely rich and active.

Duration and Sessions

We envision the following tutorial organization:

1. Introduction: data journalism and journalistic fact-checking, in early days and today
2. Use case study: analysis of a few high-visibility fact-checking tasks
3. Scientific tools and techniques which can be leveraged to help fact-checking
   - This will be organized in sub-sections by the main scientific area, as outlined above
4. Open questions, currently hot topic of work
5. Conclusions and perspectives

We propose the tutorial for half a day.

Audience
We believe the tutorial would be of interest to all conference attendees, maybe particularly so to those interested in the MisInfo track mentioned above, but also to researchers and students working in one of the related areas, as well as to practitioners which may identify common problems and learn about possible solutions.

Previous editions
The tutorial has not been conducted before.
A (very!) early embryo was presented by I. Manolescu at a Workshop on Open Data in 2013 http://www-etis.ensea.fr/WOD2013/?page_id=53) to an audience of approximately 40. The slides are available at: http://pages.saclay.inria.fr/ioana.manolescu/SLIDES/Ioana-WOD-

29052013-1.pdf and http://pages.saclay.inria.fr/ioana.manolescu/SLIDES/Ioana-WOD-29052013-2.pdf.
The four years which have passed since have however witnessed: a rapid growth of interest from the computer science community; a much better understanding from our side, of the issues involved in computational fact-checking, of the way journalists do their work, and how we could help, in particular through our collaborative project ContentCheck (mentioned above).
Thus, while our interest in the topic continued, the 2018 tutorial will be very different from the 2013 one (which we mention here only for full disclosure).

Tutorial Material
The material will be provided to the attendees. We do not foresee any copyright issues.

Equipment
We will be happy to give the tutorial using the standard equipment.

References

[ARV+16] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance Detection with Bidirectional Conditional Encoding. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Jian Su, Kevin Duh, and Xavier Carreras (Eds.). Association for Computational Linguistics, Austin, Texas, 876–885.
[BBC16] Raphael Bonaque, T. D. Cao, Bogdan Cautis, François Goasdoué, J. Letelier, Ioana Manolescu, O. Mendoza, S. Ribeiro, Xavier Tannier, and Michael Thomazo.2016. Mixed-instance querying: a lightweight integration architecture for datajournalism. PVLDB 9, 13 (2016), 1513–1516.
[BCL+17] Adnene Belfodil, Sylvie Cazalens, Philippe Lamarre, Marc Plantevit. Flash points: Discovering exceptional pairwise behaviors in vote or rating data. *ECML/PKDD*, Sep 2017, Skopje, Macedonia.
[CGY16] Barbara Catania, Giovanna Guerrini, Beyza Yaman: Context-Dependent Quality-Aware Source Selection for Live Queries on Linked Data. EDBT 2016: 716-717
[CMP11] C. Castillo, M. Mendoza, B. Poblete. 2011. Information credibility on Twitter. In Proceedings of the20th International Conference on World Wide Web, WWW '11, ACM, New York, NY, pp. 675–684.
[CMT17] Tien Duc Cao, Ioana Manolescu, Xavier Tannier: Extracting linked data from statistic spreadsheets. SBD@SIGMOD 2017
[DBS09a] Xin Luna Dong, Laure Berti-Equille, Divesh Srivastava, Integrating conflicting data: the role of source dependence. Proceedings of the VLDB Endowment, 2(1):550-561, 2009
[DBS09b] Xin Luna Dong, Laure Berti-Equille, Divesh Srivastava, Truth discovery and copying detection in a dynamic world. Proceedings of the VLDB Endowment, 2(1):562-573, 2009
[DGM05] Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL Recognising Textual Entailment Challenge. In PASCAL Challenges Workshop for Recognizing Textual Entailment.
[DS13] Dong, Xin Luna, and Srivastava, Divesh. "Big data integration." *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*. IEEE, 2013.
[FHM05] Franklin, Michael and Halevy, Alon and Maier, David: From Databases to Dataspaces: A New Abstraction for Information Management. Sigmod Records, 34(4):27--33, 2005
[GKK+13] François Goasdoué, Konstantinos Karanasos, Yannis Katsis, Julien Leblay, Ioana Manolescu, Stamatis Zampetakis: Fact checking and analyzing the web. SIGMOD Conference 2013: 997-1000
[GMG16] Guggilla, C.; Miller, T. & Gurevych, I. CNN- and LSTM-based Claim Classification in Online User Comments. *Proceedings of the 26th International Conference on Computational Linguistics (COLING 2016),* 2016

[HKN+16] Halevy, Alon and Korn, Flip and Noy, Natalya F and Olston, Christopher and Polyzotis, Neoklis and Roy, Sudip and Whang, Steven Euijong, Goods: Organizing google's datasets. Proceedings of the 2016 International Conference on Management of Data, 795--806, 2016

[HSW+14] Naeemul Hassan, Afroza Sultana, You Wu, Gensheng Zhang, Chengkai Li, Jun Yang, and Cong Yu. Data in, fact out: Automated monitoring of facts by FactWatcher. Proceedings of the Very Large Databases Conference (PVLDB), 7(13):1557–1560, 2014.

[HZA+17] Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarni, Anil Kumar Nayak, Vikas Sable, Chengkai Li, Mark Tremayne: ClaimBuster:The First-ever End-to-end Fact-checking System. 1945 - 1948.

[KTH+12] Remy Kessler, Xavier Tannier, Caroline Hagège, Véronique Moriceau, and André Bittar. Finding Salient Dates for Building Thematic Timelines. In Proceedingsof the 50th Annual Meeting of the ACL, 2012

[Leblay17] A Declarative Approach to Data-Driven Fact Checking. AAAI 2017: 147-153

[LCL17] Julien Leblay, Weiling Chen, Steven J. Lynden: Exploring the Veracity of Online Claims with BackDrop. CIKM 2017: 2491-2494

[LGM+12] Jens Lehmann, Daniel Gerber, Mohamed Morsey, and Axel-Cyrille Ngonga Ngomo. "Defacto-deep fact validation." In *International Semantic Web Conference*, pp. 312-327. Springer, Berlin, Heidelberg, 2012.

[LSD13] Amnon Lotan, Asher Stern, and Ido Dagan. 2013. TruthTeller: Annotating Predicate Truth. In Proceedings of NAACL-HLT 2013. Atlanta, USA, 752–757.

[LSW15] Xiao Ling, Sameer Singh, and Daniel Weld. 2015. Design Challenges for Entity Linking. Transactions of the Association for Computational Linguistics (TACL) 3 (2015), 315–328.

[RDD+16] Theodoros Rekatsinas, Amol Deshpande, Xin Luna Dong, Lise Getoor, Divesh Srivastava: SourceSight: Enabling Effective Source Selection. SIGMOD Conference 2016: 2157-2160

[SWH15] Wei Shen, Jianyong Wang, and Jiawei Han. 2015. Entity Linking With a Knowledge Base: Issues, Techniques, and Solutions. Transactions on Knowledge and Data Engineering (2015).

[THM15] G. Tran, E. Herder, and K. Markert. JointGraphical Models for Date Selection in Timeline Summarization.In Proceedings of the 53rd ACL, 2015

[TV16] Xavier Tannier, Frédéric Vernier. Creation, Visualization and Edition of Timelines for Journalistic Use. in *Proceedings of "Natural Language meets Journalism", workshop of the International Joint Conference on Artificial Intelligence (IJCAI 2016)*. New York, USA, July 2016.

[WAL+14] You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, and Cong Yu. Toward computational fact-checking. Proceedings of the Very Large Databases Conference (PVLDB), 7(7):589–600, 2014.