Fact Checking: Task definition and dataset construction

Andreas Vlachos
Dept. of Computer Science
University College London
London, United Kingdom
a.vlachos@cs.ucl.ac.uk

Sebastian Riedel
Dept. of Computer Science
University College London
London, United Kingdom
s.riedel@ucl.ac.uk

Abstract

In this paper we introduce the task of fact checking, i.e. the assessment of the truthfulness of a claim. The task is commonly performed manually by journalists verifying the claims made by public figures. Furthermore, ordinary citizens need to assess the truthfulness of the increasing volume of statements they consume. Thus, developing fact checking systems is likely to be of use to various members of society. We first define the task and detail the construction of a publicly available dataset using statements fact-checked by journalists available online. Then, we discuss baseline approaches for the task and the challenges that need to be addressed. Finally, we discuss how fact checking relates to mainstream natural language processing tasks and can stimulate further research.

1 Motivation

Fact checking is the task of assessing the truthfulness of claims made by public figures such as politicians, pundits, etc. It is commonly performed by journalists employed by news organisations in the process of news article creation. More recently, institutes and websites dedicated to this cause have emerged such as Full Fact[1] and PolitiFact[2] respectively. Figure 1 shows two examples of fact checked statements, together with the verdicts offered by the journalists.

Fact-checking is a time-consuming process. In assessing the first claim in Figure 1 a journalist would need to consult a variety of sources to find the average "full-time earnings" for criminal barristers. Fact checking websites commonly provide the detailed analysis (not shown in the figure) performed to support the verdict.

Automating the process of fact checking has recently been discussed in the context of computational journalism (Cohen et al., 2011; Flew et al., 2012). Inspired by the recent progress in natural language processing, databases and information retrieval, the vision is to provide journalists with tools that would allow them to perform this task automatically, or even render the articles "live" by updating them with most current data. This automation is further enabled by the increasing online availability of datasets, survey results, and reports in machine readable formats by various institutions, e.g. EUROSTAT releases detailed statistics for all European economies.[3]

Furthermore, ordinary citizens need to fact check the information provided to them. This need is intensified with the proliferation of social media such as Twitter, since the dissemination of news and information commonly circumvents the traditional news channels (Petrovic, 2013). In addition, the rise of citizen journalism (Goode, 2009) suggests that often citizens become the sources of information. Since the information provided by them is not edited or curated, automated fact checking would assist in avoiding the spreading false information.

In this paper we define the task of fact-checking. We then detail the construction of a dataset using fact-checked statements available online. Finally, we describe the challenges it poses and its relation to current research in natural language processing.

[1] http://fullfact.org

[2] http://politifact.com

[3] http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home

Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science, pages 18–22,
Baltimore, Maryland, USA, June 26, 2014. ©2014 Association for Computational Linguistics

2 Task definition

We define fact-checking to be the assignment of a truth value to a claim made in a particular context. Thus it is natural to consider it as a binary classification task. However, it is often the case that the statements are not completely true or false. For example, the verdict for the third claim in Figure 1 is MostlyTrue because some of the sources dispute it, while in the fourth example the statistics can be manipulated to support or disprove the claim as desired. Therefore it is better to consider fact-checking as an ordinal classification task (Frank and Hall, 2001), thus allowing systems to capture the nuances of the task.

The verdict by itself, even if graded, needs to be supported by an analysis (e.g., what is the systems interpretation of the statement). However, given the difficulty of carving out exactly what the correct analysis for a statement might be, we restrict the task to be a prediction problem so that we can evaluate performance automatically.

Context can be crucial in fact-checking. For example, knowing that the fourth claim of Figure 1 is made by a UK politician is necessary in order to assess it using data about this country. Furthermore, time is also important since the various comparisons usually refer to time-frames anchored at the time a claim is made.

The task is rather challenging. While some claims such as the one about Crimea can be fact-checked by extracting relations from WikiPedia, the verdict often hinges on interpreting relatively fine points, e.g. the last claim refers to a particular definition of income. Journalists also check multiple sources in producing their verdicts, as in the case of the third claim. Interestingly, they also consider multiple interpretations of the data; e.g. in the last claim is assessed as HalfTrue since different but reasonable interpretations of the same data lead to different conclusions.

We consider all of the aspects mentioned (time, speaker, multiple sources and interpretations) as part of the task of fact checking. However, we want to restrict the task to statements that can be fact-checked objectively, which is not always true for the statements assessed by journalists. Therefore, we do not consider statements such as "New Labour promised social improvement but delivered a collapse in social mobility" to be part to the task since there are no universal definitions of "social improvement" and "social mobility".[4]

Claim *(by Minister Shailesh Vara)*

"The average criminal bar barrister working full-time is earning some £84,000."

Verdict: False *(by Channel 4 Fact Check)*

The figures the Ministry of Justice have stressed this week seem decidedly dodgy. Even if you do want to use the figures, once you take away the many overheads self-employed advocates have to pay you are left with a middling sum of money.

Claim *(by U.S. Rep. Mike Rogers)*

"Crimea was part of Russia until 1954, when it was given to the Soviet Republic of the Ukraine."

Verdict: True *(by Politifact)*

Rogers said Crimea belonged to Russia until 1954, when Khrushchev gave the land to Ukraine, then a Soviet republic.

Claim *(by President Barack Obama)*

"For the first time in over a decade, business leaders around the world have declared that China is no longer the world's No. 1 place to invest; America is."

Verdict: MostlyTrue *(by Politifact)*

The president is accurate by citing one particular study, and that study did ask business leaders what they thought about investing in the United States. A broader look at other rankings doesn't make the United States seem like such a powerhouse, even if it does still best China in some lists.

Claim *(by Chancellor George Osborne)*

"Real household disposable income is rising."

Verdict: HalfTrue *(by Channel 4 Fact Check)*

RHDI did grow in latest period we know about (the second quarter of 2013), making Mr Osborne arguably right to say that it is rising as we speak. But over the last two quarters we know about, income was down 0.1 per cent. If you want to compare the latest four quarters of data with the previous four, there was a fall in household income, making the chancellor wrong. But if you compare the latest full year of results, 2012, with 2011, income is up and he's right again.

Figure 1: Fact-checked statements.

[4] http://blogs.channel4.com/factcheck/factcheck-social-mobility-collapsed/

3 Dataset construction

In order to construct a dataset to develop and evaluate approaches to fact checking, we first surveyed popular fact checking websites. We decided to consider statements from two of them, the fact checking blog of Channel 4[5] and the Truth-O-Meter from PolitiFact.[6] Both websites have large archives of fact-checked statements (more than 1,000 statements each), they cover a wide range of prevalent issues of U.K. and U.S. public life, and they provide detailed verdicts with fine-grained labels such as MOSTLYFALSE and HALFTRUE.

We examined recent fact-checks from each website at the time of writing. For each statement, apart from the statement itself, we recorded the date it was made, the speaker, the label of the verdict and the URL. As the two websites use different labelling schemes, we aligned the labels of the verdicts to a five-point scale: TRUE, MOSTLYTRUE, HALFTRUE, MOSTLYFALSE and FALSE. The speakers included, apart from public figures, associations such as the American Beverage Association, activists, even viral FaceBook posts submitted by the public.

We then decided which of the statements should be considered for the task proposed. As discussed in the previous section we want to avoid statements that cannot be assessed objectively. Following this, we deemed unsuitable statements:

- assessing causal relations, e.g. whether a statistic should be attributed to a particular law
- concerning the future, e.g. speculations involving oil prices
- not concerning facts, e.g. whether a politician is supporting certain policies

For the statements that were considered suitable, we also collected the sources used by the journalists in the analysis provided for the verdict. Common sources include tables with statistics and reports from governments, think tanks and other organisations, available online. Automatic identification of the sources needed to fact check a statement is an important stage in the process, which is potentially useful in its own right in the context of assisting journalists in a semi-automated fact-checking approach Cohen et al. (2011). Sometimes the verdicts relied on data that were not available online such personal communications; statements whose verdict relied on such data were also deemed unsuitable for the task.

As mentioned earlier, the verdicts on the websites are accompanied by lengthy analyses. While such analyses could be useful annotation for intermediate stages of the task — e.g. we could use it as supervision to learn how to combine the information extracted from the various sources into a verdict — we noticed that the language used in them is indicative of the verdict.[7] Thus we decided not to include them in the dataset, as it would enable tackling part of the task as sentiment analysis. Out of the 221 fact-checked statements examined, we judged 106 as suitable. The dataset collected including our suitability judgements is publicly available[8] and we are working on extending it so that it can support the development and the automatic evaluation of fact checking approaches.

4 Baseline approaches

As discussed in Section 2, we consider fact checking as an ordinal classification task. Thus, in theory it would be possible to tackle it as a supervised classification task using algorithms that learn from statements annotated with the verdict labels. However this is unlikely to be successful, since statements such as the ones verified by journalists do not contain the world knowledge and the temporal and spatial context needed for this purpose.

A different approach would be to match statements to ones already fact-checked by journalists and return the label in a K-nearest neighbour fashion.[9] Thus the task is reduced to assessing the semantic similarity between statements, which was explored in a recent shared task (Agirre et al., 2013). An obvious shortcoming of this approach is that it cannot be applied to new claims that have not been fact-checked, thus it can only be used to detect repetitions and paraphrases of false claims.

A possible mechanism to extend the coverage of such an approach to novel statements is to assume that some large text collection is the source of all true statements. For example, Wikipedia is likely

16444

[5] http://blogs.channel4.com/factcheck/

[6] http://www.politifact.com/truth-o-meter/statements/

[7] E.g. part of the analysis of the first claim in Figure 1 reads: "the full-time figure has the handy effect of stripping out the very lowest earners and bumping up the average".

[8] https://sites.google.com/site/andreasvlachos/resources

[9] The Truth-Teller by Washington Post (http://truthteller.washingtonpost.com/) follows this approach.

to contain a statement that would match the second claim in Figure 1. However, it would still be unable to tackle the other claims mentioned, since they require calculations based on the data.

5 Discussion

The main drawback of the baseline approaches mentioned (aside from their potential coverage) is the lack of interpretability of their verdicts, also referred to as algorithmic accountability (Diakopoulos, 2014). While it is possible for a natural language processing expert to inspect aspects of the prediction such as feature weights, this tends to become harder as the approaches become more sophisticated. Ultimately, the user of a fact checking system would trust a verdict only if it is accompanied by an analysis similar to the one provided by the journalists. This desideratum is present in other tasks such as the recently proposed science test question answering (Clark et al., 2013).

Cohen et al. (2011) propose that fact checking is about asking the right questions. These questions might be database queries, requests for information to be extracted from textual resources, etc. For example, in checking the last claim in Figure 1 a critical reader would like to know what are the possible interpretations of "real household disposable income" and what the calculations might be for other reasonable time spans.

The manual fact checking process suggests an approach that is more likely to give an interpretable analysis and would decompose the task into the following stages:

1. extract statements to be fact-checked
2. construct appropriate questions
3. obtain the answers from relevant sources
4. reach a verdict using these answers

The stages of this architecture can be mapped to tasks well-explored in the natural language processing community. Statement extraction could be tackled as a sentence classification problem, following approaches similar to those proposed for speculation detection (Farkas et al., 2010) and veridicality assessment (de Marneffe et al., 2012). Furthermore, obtaining answers to questions from databases is a task typically addressed in the context of semantic parsing research, while obtaining such answers from textual sources is usually considered in the context of information extraction. Finally, the compilation of the answers into a verdict could be considered as a form of logic-based textual entailment (Bos and Markert, 2005).

However, the fact-checking stages described include a novel task, namely question construction for a given statement. This task is likely to rely on semantic parsing of the statement followed by restructuring of the logical form generated. Since question construction is a rather uncommon task, it is likely to require human supervision, which could possibly be obtained via crowdsourcing. Furthermore, the open-domain nature of fact checking places greater demands on the established tasks of information extraction and semantic parsing. Thus, fact-checking is likely to stimulate research in these tasks on methods that do not require domain-specific supervision (Riedel et al., 2013) and are able to adapt to new information requests (Kwiatkowski et al., 2013).

Fact-checking is related to the tasks of textual entailment (Dagan et al., 2006) and machine comprehension (Richardson et al., 2013), with the difference that the text which should be used to predict the entailment of the hypothesis or the correct answer respectively is not provided in the input. Instead, systems need to locate the sources needed to predict the verdict label as part of the task. Furthermore, by defining the task in the context of real-world journalism we are able to obtain labeled statements at no annotation cost, apart from the assessment of their suitability for the task.

6 Conclusions

In this paper we introduced the task of fact checking and detailed the construction of a dataset using statements fact-checked by journalists available online. In addition, we discussed baseline approaches that could be applied to perform the task and the challenges that need to be addressed.

Apart from being a challenging testbed to stimulate progress in natural language processing, research in fact checking is likely to inhibit the intentional or unintentional dissemination of false information. Even an approach that would return the sources related to a statement could be very helpful to journalists as well as other critical readers in a semi-automated fact checking approach.

Acknowledgments

The authors would like to thank the members of the Machine Reading lab for useful discussions

and their help in compiling the dataset.

References

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *sem 2013 shared task: Semantic textual similarity. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics and the Shared Task: Semantic Textual Similarity*, pages 32–43, Atlanta, GA.

Johan Bos and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP 2005)*, pages 628–635.

Peter Clark, Philip Harrison, and Niranjan Balasubramanian. 2013. A study of the knowledge base requirements for passing an elementary science test. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, pages 37–42.

Sarah Cohen, Chengkai Li, Jun Yang, and Cong Yu. 2011. Computational journalism: A call to arms to database researchers. In *Proceedings of the Conference on Innovative Data Systems Research*, volume 2011, pages 148–151.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment*, pages 177–190.

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2012. Did it happen? the pragmatic complexity of veridicality assessment. *Computational Linguistics*, 38(2):301–333, June.

Nick Diakopoulos. 2014. Algorithmic accountability reporting: On the investigation of black boxes. Technical report, Tow Center for Digital Journalism.

Richard Farkas, Veronika Vincze, Gyorgy Mora, Janos Csirik, and Gyorgy Szarvas. 2010. The CoNLL 2010 Shared Task: Learning to Detect Hedges and their Scope in Natural Language Text. In *Proceedings of the CoNLL 2010 Shared Task*.

Terry Flew, Anna Daniel, and Christina L. Spurgeon. 2012. The promise of computational journalism. In *Proceedings of the Australian and New Zealand Communication Association Conference*, pages 1–19.

Eibe Frank and Mark Hall. 2001. A simple approach to ordinal classification. In *Proceedings of the 12th European Conference on Machine Learning*, pages 145–156.

Luke Goode. 2009. Social news, citizen journalism and democracy. *New Media & Society*, 11(8):1287–1305.

Tom Kwiatkowski, Eunsol Choi, Yoav Artzi, and Luke Zettlemoyer. 2013. Scaling semantic parsers with on-the-fly ontology matching. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1556, Seattle, WA.

Sasa Petrovic. 2013. *Real-time event detection in massive streams*. Ph.D. thesis, School of Informatics, University of Edinburgh.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, WA.

Sebastian Riedel, Limin Yao, Benjamin M. Marlin, and Andrew McCallum. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, GA.