Do “Future Work” sections have a purpose? Citation links and entailment for global scientometric questions

Simone Teufel

Computer Laboratory, University of Cambridge, and Computer Science Department, Tokyo Institute of Technology

Abstract. Which tasks in digital libraries might be interesting and addressable for us as a community now in the near future, given the recent developments in NLP? This paper attempts to make a guess. (Instead of actually even attempting to answer the question in its title.) What would we need to do today if we, at some point in the future, wanted to be able to answer this question objectively, quantifiably, and on a large text base? Many similar questions in the same vein exist, such as where the research of a field as a whole is headed, where the currently most contested research issues in a field are, which new ideas emerged in the past 5 years, and which ones of these are game-changers. I will argue that we might well be able to offer support for the answering of such questions sometime in the future, particularly if we are willing to turn our attention to entailment and inference in scientific writing.

1 Introduction

Something always irks me when I read “future work” sections; be it the “future work” sections of papers I review, those of papers I read for my research, or the ones I have to compose when writing my own papers. Questions such as – what exactly is the purpose of these sections, and what is the status of the ideas in these sections? How should I interpret these descriptions in other papers, and how should I write my own?

In terms of communicative function, what a “future work” section is supposed to accomplish is less clear-cut than, for instance, a “methods” section. Ideas we express there don’t really “count” in terms of the contributions of the paper: the research ideas we present there are hypothetical (they are not yet done), so no real contribution can be claimed for them. Lacking a potential reward, why would we offer up our unprotected secret research plans to anybody who might want to snatch them? On the other hand, social convention requires that we write *something*; we can’t just leave the space empty. True, occasionally one has some ongoing work that one urgently wants to tell the community about, or a particularly clever idea that fits neatly. But I suspect that many of the ideas I read about are mechanically written possible avenues for future research, which get forgotten almost as soon as the paper is published. Maybe they get taken up by somebody – more likely not. (I know this because I sometimes do this myself.)

But "future work" sections could also in principle be like a market for ideas, a notice board where we announce our true intentions, and where we compete with our readers – just who manages to beat the time and bring out the next paper on this hot future topic? Perhaps the main inspiration for a large proportion of papers ever written was indeed such an open research question previously posed in a paper the authors read?

I wished that an enterprising social scientist, bibliometrist or historian of science did this research, and I also wished he or she would (be able to) use the large data repositories we now have at our fingertips when doing bibliographic searches. I would personally like a quantifiably answer, which is precise and supported by much textual historical evidence. Could this possibly be an application of large-scale digital libraries in combination with natural language technology? Maybe not quite yet, but thinking about what we would need in order to answer this question could provide some near future goals for NLP.

The short answer to what we would need to do is to match up descriptions of planned research in an earlier set of papers, with research contributions of papers later in time (either by the same or by different authors). The detailed answer is more complicated and concerns technologies such as paraphrase detection, citation link processing, citation block detection, entailment detection and many more.

1.1 A Simple Example

Suppose we find the following sentence in the "future work" section of a paper.

(1) *We intend to investigate more sophisticated ways of document representation and of extracting a citation's context.*

We then start a search for this task, and given that the communicative purpose of a scientific paper (its knowledge claim) is generally quite clearly marked, let's say we find a paper (by the same authors) in the future of the first paper (say, a year later), with the following title:

(2) *Context Matters: Towards Extracting a Citations Context Using Linguistic Features*

We can't help concluding that the original intention of the authors, stated in sentence (1) above, must have been real, as it was obviously followed up by exactly the kind of research predicted. Manually finding examples where *other researchers* take up a suggestion from a previous "future work" section is a bit harder to do. Of course, in many cases we might never find a good match.

It is clear that a search engine that could provide a social scientist with objective, hard data on how many times a plausible match occurs would have to be pretty intelligent. It is definitely addressing a text understanding problem. (My definition of "text understanding" is "any process that obtains *new* knowledge, i.e., something that is true and relevant but that isn't explicitly stated in the text). My guess is, however, that such a system wouldn't have to be *impossibly* intelligent, given today's NLP technology.

1.2 Paraphrases, pragmatic effects, and concreteness

The first component we would need is very robust paraphrase detection. Research in paraphrase detection is long-standing, with many successful approaches [3, 7, 12, 31]. But we would need capability going beyond paraphrase detection. In particular, we will need to draw inferences. Understanding how human inference works, and sometimes being able to automate some of these steps, is an important part of text understanding, and thus of artificial intelligence–a worthwhile exercise in its own right. For the current task at hand, being able to perform inference would also happen to be extremely useful.

The closest we have come as a field to a shared inference task is the RTE [11], and its precursor, FRACAS [10]. In these tasks, systems have to verify or reject a proposed logical entailment hypothesis between two sentences. While FRACAS works with hand-crafted sentence pairs and concentrates on entailments that follow from the manipulation of negation, quantifiers and scope, RTE uses naturally occurring text. The inferences in RTE generally also require world knowledge (this is the case for FRACAS to a much lower degree), and because world knowledge is often not shared exactly across humans, the notion of entailment becomes somewhat *defeasible*; ie., instead of "strictly logical" entailment, a weaker notion of "plausible" entailment (that would generally be accepted by most humans) is used.

This has more recently been complemented by *visual entailment* tasks, such as the one based on image captions that underlies the SNLI, the Stanford Natural Language Inference corpus [5]. This is a large-scale task consisting of around 570,000 sentence pairs. The possible hypotheses (entailed ones, non-entailed ones, and possibly-entailed ones) were elicited from naive humans who did not see the original image, but only had access to the caption as the basis on which to form their hypotheses. Of course, such a task definition excludes cognitive, communicative and other abstract action, events and objects, which are an essential part of the inference necessary for understanding real-world text such as scientific papers, but it does allow for the training of supervised machine learning algorithms such as neural networks which require vast training material.

Recognising and processing presuppositions, and implicatures in general, can also help immensely. Apart from knowing about what is logically entailed in a text, knowing what is *implicated* is another important mosaic stone in the inference puzzle. Implicatures are defined as those statements about the world that are assumed to be true by the speaker and transmitted "along with" their literal message, without being explicitly stated.

(3) a. *Miller et al. did not manage to verify whether saturation was reached.*
b. *Miller et al. did not verify whether saturation was reached.*

In contrast to entailment, implicatures cannot be cancelled by negation. If I state sentence (3-a) above, we understand that Miller et al. attempted the verification, but that the test was inconclusive. If I state sentence (3-b) above, I trigger a very different meaning, namely that they didn't want to verify, and probably didn't even run the test. Please also note that the presupposition "they had the intention of verifying" survives even if we turn the negative sentence into a positive one.

This is a remarkable difference, in that sentences (3-a) and (3-b) are truth-conditionally equivalent; the only thing that distinguishes them is the verb *manage to*. Presuppositions are distinguished from general implicatures in that they are lexically trigged.

(4) a. *Miller attempted to model how X works.*
b. *Miller modelled how X works.*

Another example concerns conversational implicatures. Stating sentence (4-a) above conversationally implicates that Miller et al. didn't manage to satisfactorily model how X works. (If they had, in the author's opinion, then sentence (4-b) would have been more appropriate. This follows from the Gricean maxims [16], according to which one should always state the strongest relevant and true statement one possibly can. "modelling" (i.e., attempting to model and then succeeding) is stronger than simply "attempting to model".

There are few works that model pragmatic reasoning effects computationally on a large scale, but some resources such as presupposition trigger lexicons have been created, and some automatic research exists on detecting presuppositions and implicatures beween pairs of verbs, e.g. [29]. In Cambridge, there is some recent work on determining how the inference involved in interpreting "let alone" sentences can be automated [27], which also involves decisions on possible presupposition links between two statements.

Overall, I believe that such works can contribute towards the question in the title of this paper – finding links between people's proclaimed intentions of doing something (as stated in the "future work" sections), and the reports of actually having done that thing (in a research paper). They can potentially also help find similarities between methods used in different papers, and determine what the exact difference might be in cases where a contrast between some methods exists but where the nature of the difference is not spelled out by the authors.

Scientific writing, like all texts, contain many implicatures. According to the Gricean Maxims, we would be insulting the reader's intelligence (or sound pompous) if we tell them more than they need to know to make the inference themselves, and the search space for possible inferences is very, very large. In my opinion, any work on automatic inference should therefore start with the clear-cut, small-step cases, such as presuppositions and implicatures, where there is normally very high human agreement that these statements have objectively been asserted into the discourse, and what the implicature is. This is opposed to other types of inference, where much more world knowledge is needed to make the inference.

Finally, I will propose that judging the "abstractness" of a "future work" item would also be useful. This could serve to distinguish the following pair of sentences:

(5) a. In future work we intend to evaluate the algorithm as part of a dialogue understanding system on state of the art benchmarks.
b. . . . it is not known exactly how well the model will perform in the real world. Future work will examine installing models in real world applications.

The first avenue for future work (sentence (5-a)) is quite concrete, and the chances are high that the researchers, having already implemented their algorithm, might conceivably next perform the rather concrete action of running an evaluation. In sentence (5-b), the future work suggested sounds abstract and vague, and like work of a different kind altogether, for which the research team probably are neither equipped, nor have any real intention of doing. Research in metaphor classification has contributed some methods for addressing the task of judging the abstractness of phrases [30, 26]

1.3 Global Scientometric Questions

Predicting which "future work" suggestion is taken up in later work is an exciting task. An additional question is that if some such suggestion does get taken up, whether it is by the authors of the original paper or by somebody else. Given an acceptably precision matching procedure, a natural gold standard for the prediction task offers itself, because in citation-based research we can manipulate the time scale of papers we take into account, in order to simulate a "future" that is to be predicted.

There are various related scientometric questions we might ask about the development of a field, such as the task of describing the emergence and development of a scientific field [4, 19], determining schools of thought [22, 1], finding the occurrence of scientific revolutions and paradigm shifts [20, 13], identifying the scientific areas where the most innovation currently occurs [9, 6], and detecting the emergence of scientific ideas [21, 23]. These questions have traditionally been answered manually or purely quantitatively, and more recently with the help of natural language processing technology such as automatic sentiment detection and citation function classification [25, 15, 24, 28, 18, 2, 14, 8, 17]. My argument here is that we need not stop there. In my opinion, these existing methods could gainfully be complemented with computational linguistics research studying textual entailment, text understanding and pragmatics of scientific writing.

References

1. Allen, B.: Referring to schools of thought: An example of symbolic citations. Social Studies of Science 27(6), 937–949 (1997)

2. Athar, A., Teufel, S.: Detection of implicit citations for sentiment detection. In: Proceedings of ACL-12 Workshop on Discovering Structure in Scholarly Discourse. Jeju Island, South Korea (2012)
3. Barzilay, R., Lee, L.: Learning to paraphrase: an unsupervised approach using multiple-sequence alignment. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1. pp. 16–23. Association for Computational Linguistics (2003)
4. Bettencourt, L.A., Ulwick, A.W.: The customer-centered innovation map. Harvard Business Review 86(5), 109 (2008)
5. Bowman, S.R., Angeli, G., Potts, C., Manning, C.D.: A large annotated corpus for learning natural language inference. arXiv preprint arXiv:1508.05326 (2015)
6. Boyack, K.W.: Thesaurus-based methods for mapping contents of publication sets. Scientometrics 111(2), 1141–1155 (2017)
7. Callison-Burch, C.: Syntactic constraints on paraphrases extracted from parallel corpora. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 196–205. Association for Computational Linguistics (2008)
8. Catalini, C., Lacetera, N., Oettl, A.: The incidence and role of negative citations in science. Proceedings of the National Academy of Sciences 112(45), 13823–13826 (2015)
9. Chen, C., Ibekwe-SanJuan, F., Hou, J.: The structure and dynamics of cocitation clusters: A multiple-perspective cocitation analysis. Journal of the Association for Information Science and Technology 61(7), 1386–1409 (2010)
10. Cooper, R., Crouch, D., Van Eijck, J., Fox, C., Van Genabith, J., Jaspars, J., Kamp, H., Milward, D., Pinkal, M., Poesio, M., et al.: Using the framework. Tech. rep., Technical Report LRE 62-051 D-16, The FraCaS Consortium (1996)
11. Dagan, I., Glickman, O., Magnini, B.: The pascal recognising textual entailment challenge. In: Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment, pp. 177–190. Springer (2006)
12. Das, D., Smith, N.A.: Paraphrase identification as probabilistic quasi-synchronous recognition. In: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1. pp. 468–476. Association for Computational Linguistics (2009)
13. De Langhe, R.: Towards the discovery of scientific revolutions in scientometric data. Scientometrics 110(1), 505–519 (2017)
14. Ding, Y., Zhang, G., Chambers, T., Song, M., Wang, X., Zha, C.: Content-based citation analysis: The next generation of citation analysis. Journal of the Association for Information Science and Technology 65 (2014)
15. Garzone, M., Mercer, R.E.: Towards an automated citation classifier. In: Proceedings of the 13th Biennial Conference of the CSCI/SCEIO (AI-2000). pp. 337–346 (2000)
16. Grice, H.P., Cole, P., Morgan, J., et al.: Logic and conversation. 1975 pp. 41–58 (1975)
17. Jha, R., Jbara, A.A., Qazvinian, V., Radev, D.R.: Nlp-driven citation analysis for scientometrics. Natural Language Engineering 23(1), 93–130 (2017)
18. Kaplan, D., Iida, R., Tokunaga, T.: Automatic extraction of citation contexts for research paper summarization: A coreference-chain based approach. In: Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries. pp. 88–95. Association for Computational Linguistics (2009)

19. Kiss, I.Z., Broom, M., Craze, P.G., Rafols, I.: Can epidemic models describe the diffusion of topics across disciplines? Journal of Informetrics 4(1), 74–82 (2010)
20. Kuhn, T.S.: The Structure of Scientific Revolutions, 2nd enl. ed. University of Chicago Press (1970)
21. Kuhn, T., Perc, M., Helbing, D.: Inheritance patterns in citation networks reveal scientific memes. Physical Review X 4(4), 041036 (2014)
22. McCain, K.W.: Cocited author mapping as a valid representation of intellectual structure. Journal of the American society for information science 37(3), 111 (1986)
23. McKeown, K., Daume, H., Chaturvedi, S., Paparrizos, J., Thadani, K., Barrio, P., Biran, O., Bothe, S., Collins, M., Fleischmann, K.R., et al.: Predicting the impact of scientific concepts using full-text features. Journal of the Association for Information Science and Technology 67(11), 2684–2696 (2016)
24. Nakov, P., Schwarz, A., Hearst, M.: Citances: Citation sentences for semantic analysis of bioscience text. In: SIGIR'04 Workshop on Search and Discovery in Bioinformatics (2004)
25. Nanba, H., Okumura, M.: Towards multi-paper summarization using reference information. In: Proceedings of the XXth International Joint Conference on Artificial Intelligence (IJCAI-99). pp. 926–931 (1999)
26. Neuman, Y., Assaf, D., Cohen, Y., Last, M., Argamon, S., Howard, N., Frieder, O.: Metaphor identification in large texts corpora. PloS one 8(4), e62343 (2013)
27. Razuvayevskaya, O., Teufel, S.: Finding enthymemes in real-world texts: a feasibility study. Argument and Computation (2017)
28. Teufel, S., Siddharthan, A., Tidhar, D.: Automatic classification of citation function. In: Proceedings of EMNLP-06 (2006)
29. Tremper, G., Frank, A.: A discriminative analysis of fine-grained semantic relations including presupposition: Annotation and classification. Dialogue & Discourse 4(2), 282–322 (2013)
30. Turney, P.D., Neuman, Y., Assaf, D., Cohen, Y.: Literal and metaphorical sense identification through concrete and abstract context. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 680–690. Association for Computational Linguistics (2011)
31. Xu, W., Ritter, A., Grishman, R.: Gathering and generating paraphrases from twitter with application to normalization. In: Proceedings of the Sixth Workshop on Building and Using Comparable Corpora. pp. 121–128 (2013)