Lifetime Achievement Award

Kathy McKeown Interviews Bonnie Webber

Bonnie Webber
University of Edinburgh
School of Informatics
bonnie.webber@ed.ac.uk

Because the 2020 ACL Lifetime Achievement Award presentation could not be done in person, we replaced the usual LTA talk with an interview between Professor Kathy McKeown (Columbia University) and the recipient, Bonnie Webber. The following is an edited version of the interview, with added citations.

Bonnie Webber: Let me start by thanking you, Kathy, for agreeing to interview me. I think you were the first Ph.D. student I met when I joined the faculty at Penn after finishing my own Ph.D. You were Aravind Joshi's Ph.D. student at the time, working on generating answers to definition questions. I believe your Ph.D. thesis then became the first book published in the Cambridge University Press series on Natural Language Processing, of which Aravind was the editor-in-chief.
Kathy McKeown: Well, Bonnie, thanks so much for asking me to do this. I really will enjoy it.

What I remember of you from that time was your arrival in the department like a breath of fresh air. You were very animated, always vocal. And to a young woman like myself at that time, you were an amazing role model. So I'm happy to have this chance to ask you more questions than I could for my keynote address.

You've frequently asserted that we shouldn't develop computational models based simply on standard assumptions about a linguistic phenomenon, but rather that we should first try to characterize it as fully as possible and then see which of these can be covered computationally. Given that current approaches in NLP don't focus on subtle linguistic differences and what a model does or doesn't cover, but instead worry about scale, do you think your approach to research still has something to contribute?
Bonnie: Well, I certainly hope so! Let me answer the question in terms of standard metrics used in Natural Language Processing—in particular, F-score, Precision, and Recall. Although F-score has mostly been used as a basis for "League Table" ratings, I would contend that there is much to be learned from its components—Precision and Recall. I always encourage my students to look at them very carefully.

Let me elaborate. Precision quantifies incorrect positive predictions—cases where a system has generalized incorrectly about what belongs to the class of interest, saying more things belong to it than actually do. Recall, on the other hand, quantifies incorrect negative predictions—cases where a system has generalized incorrectly about

https://doi.org/10.1162/COLI_a_00393

what doesn't belong to the class of interest, omitting things from the class that should be there.

Now, there are at least two possible reasons for low Precision and/or low Recall: Either the system itself is unable to learn from the data because it's too puny, or there's something wrong with the data. What doesn't surprise me, having spent years creating annotated corpora, is how easy it is for the latter to be the case. That is, errors in Precision and Recall show that there's more to a phenomenon than we thought.

So what I'm trying to say is that there is a lot of linguistics that's not a solved problem, and that errors in Recall and Precision should encourage people to try to better understand and characterize the phenomenon. Even if a system looks good, in terms of its True Positives and True Negatives, many of the tokens it labels correctly may just be minor variations of one another. On the other hand, looking at what has been incorrectly rejected or accepted can really tell you a lot about a phenomenon and lead to a better result on the next iteration. So, yes, I think an analytic approach to research really does have something to contribute.

Kathy: Could you give an example of this that you've come across personally?

Bonnie: Oh yes. I'd be happy to give one. When we started creating the Penn Discourse Tree Bank—that is, when we started annotating discourse coherence (relations that are signaled either explicitly with some kind of connective or implicitly by inference), we followed the standard practice of having annotators label each relation with the sense that was taken to hold between its arguments. For example, the situation described by one argument could be taken to be the result of the situation described by the other argument.

But then we had examples like the following (Rohde et al. 2017, 2018), which is made up, rather than being taken from *The Wall Street Journal* corpus:

(1) It's too far to walk. Let's take the bus.

Here, there is no explicit signal of any relation that holds between them. But people understand two distinct senses as holding, both at the same time. One can be made explicit with the connective *so*. That is, the suggestion made in the second sentence is a result of the situation described in the first sentence:

(2) It's too far to walk, so let's take the bus.

The other sense can be made explicit with the connective *instead*. That is, people understand that taking the bus is meant to be a substitute for walking:

(3) It's too far to walk. Instead, let's take the bus.

Putting them together shows that both senses hold simultaneously.

(4) It's too far to walk. So instead, let's take the bus.

Back in 2006, we modified the tool we used in annotating the PDTB to allow annotaters to record either one or two senses as holding between the arguments to a relation. However, we didn't either insist that they record multiple senses or even remind them that they could. So there weren't many tokens annotated with multiple senses in the PDTB-2 (Prasad, Webber, and Joshi 2014), and they were ignored in nearly all of the work done on learning to annotate discourse relations automatically (Lin, Ng, and Kan 2012; Xue et al. 2015, 2016).

Anyway, we moved to allowing annotators to label relations with either one or two senses from noticing the kind of disagreements they had with each other and from seeing for ourselves that the labels they were assigning all basically seemed correct. I'm now hoping that people will start taking seriously the concept of multiple sense labels and distinguishing cases where multiple relations hold from when only one or the other holds. This should be easier, given the larger number of discourse relations annotated with multiple senses in the enlarged PDTB-3 corpus (Webber et al. 2019).

Kathy: When you started in the field in the 1970s and the 1980s, any research that didn't involve syntax or parsing was taken to be on the lunatic fringe of NLP. More recently, we've seen that syntax and parsing seem to be receding and are no longer seen as relevant. They're being replaced by neural models that reject explicit notions of syntax and parsing, and the same is true of semantics. So, do you think that neural models will also successfully encode discourse, with discourse theory and discourse processing also becoming irrelevant?

Bonnie: That's a good question. I'm going to frame my answer first, in terms of discourse structure and the knowledge and reasoning needed to handle discourse phenomena such as coreference and discourse coherence relations, and second, in terms of the way that a lot of material is left implicit in text or speech because it can be gleaned from context or inferred.

With respect to discourse structure, discourse generally has some kind of macro structure that often reflects the genre of the text or speech being analyzed and the kind of information being conveyed. Examples include answers to definition questions where, as you detail in your Ph.D. thesis (McKeown 1985), answering a question like *What is an X?* is often answered by giving a statement identifying what an X is and then giving one or more illustrations of X, and describing the attributes of an X.

Of course, by now everybody knows that news articles have a macro structure consisting of a title and then an optional strap line, followed by text in an inverted pyramid structure (Stede 2012; Webber, Egg, and Kordoni 2012). Similarly, biomedical research papers (and more recently, their abstracts) are written with a macro structure that comprises a statement of the objectives of the work, the methods used in the work, the results that were obtained, and then a discussion of the results with respect to the objectives (Lin et al. 2006; Mizuta et al. 2006; Hirohata et al. 2008; Agarwal and Yu 2009; Cohen et al. 2010; Liakata et al. 2010).

Now, while computational linguistics papers do not share the same structure as biomedical research papers, it's likely from a recognition perspective that neural models will indeed be effective at recognizing this macro level discourse structure, because the features needed to recognize it are very local and there's nothing really deep about it. Also, there's a lot of available text. So we just need researchers working on the problem.

By the way, listeners might be interested to learn how news articles came to adopt the "inverted pyramid" structure: Its adoption goes back to the days when news began to be transmitted quickly across the United States by the new technology of the telegraph in the 1850s.[1] If the transmission dropped, the "inverted pyramid" ensured that the most important information would not have been lost.

[1] https://www.poynter.org/reporting-editing/2003/birth-of-the-inverted-pyramid-a-child-of-technology-commerce-and-history/.

More recently, neural models are beginning to be used to identify and exploit the reasoning, world knowledge, and pragmatics needed in order to understand text—in particular, to handle the kinds of ambiguity, under-specification and information conveyed implicitly through context or inference. I do believe there will be some success at encoding discourse, but I don't think it means that discourse theory and discourse processing will become irrelevant.

As you know, the standard illustration of the need for world knowledge and reasoning in resolving ambiguous pronouns comes from the minimal pair that Terry Winograd introduced in his 1970 Ph.D. thesis (Winograd 1972). This is the famous example contrasting the pair of sentences

(5) a. The City Council refused the women a permit because they feared violence.
b. The City Council refused the women a permit because they advocated violence.

In the case of fearing violence, readers take the pronoun *they* to refer to the City Council (whom readers take to be a bunch of wusses), while in the case of advocating violence, readers take *they* to refer to the women. Now, nearly 40 years after Terry introduced these examples in this thesis, in 2011, the examples became the basis for what came to be called the *Winograd Schema Challenge*.[2] The challenge required systems to resolve such ambiguous pronouns to one of two given antecedents.

Eight years later, in 2019, a team from Oxford, Imperial, and the Alan Turing Institute (Kocijan et al. 2019b) showed that fine-tuning a BERT language model with several million artificial Winograd schema sentences extracted from the *Wikipedia* could enable that model to correctly resolve over 75% of the examples in the Challenge, and also improve the performance of systems in resolving more prosaic examples of personal pronoun anaphors (Kocijan et al. 2019a).

Now, although this is very impressive and something we should learn from, it is unclear how performance can further improve, given that it reflects general patterns of personal pronoun use and not any specific understanding of coreference. For example, as much cleverness would be involved in creating versions of *WikiCrem* to use in training systems to handle event anaphors, bridging anaphors, and other forms of ambiguity, under-specification, and information conveyed implicitly. As such, there is still a lot for current and future NLP researchers to do—and for languages other than English as well!

Kathy: So as my final question, I want to ask whether you feel that discourse or any other field in which you've worked has changed over the course of your career.

Bonnie: Well, I'm always surprised to be reminded how long my career has been! As an M.Sc. student at Harvard, I was fortunate to be offered a part-time job at Bolt Beranek and Newman (BBN), working on question answering, as part of Bill Woods's LUNAR system (Woods, Kaplan, and Nash-Webber 1972; Woods 1978). My responsibility was to enable LUNAR to handle more types of questions, and to support more ways of asking any particular question. My Ph.D. thesis, which I completed while at BBN, characterized the kinds of discourse anaphors found in questions posed to LUNAR, and what kind of entities had to be available for those anaphors to be resolved (Webber 1979, 1982).

When I went to Penn in 1978, Aravind Joshi was also carrying out and supervising work on question answering, among all the many other things that he was interested in, such as code switching and formal languages and the complexity of Natural Language

[2] http://commonsensereasoning.org/winograd.html.

grammar. Besides your own work, Aravind's supervision here included Jerry Kaplan's work on correcting factual misconceptions underlying a question (Kaplan 1982; Joshi, Webber, and Weischedel 1984; Webber 1986), and Eric Mays' work on possible versus impossible changes in dynamic databases (Mays, Joshi, and Webber 1982; Mays 1984; Webber 1986). In both cases, it was assumed that directly answering a question that had some underlying misconception would mislead the questioner and prolong the misconception. This wasn't the only research on what, at the time, we called *cooperative question answering* (e.g., Cheikes and Webber 1988; Cheikes 1991), a label adopted by other researchers as well, including Farah Benamara and Patrick Saint-Dizer.

However, there was a big change in question answering in the 1990s, when NIST instituted question answering tracks as part of its annual Text Retrieval and Evaluation Conference (TREC). Systems competing in these tracks took an information retrieval approach to question answering—searching free text for answers to factoid questions, which have a single answer, or for answers to list questions or definition questions (cf. `https://aclweb.org/aclwiki/Question_Answering_(State_of_the_art)`). This was then followed by a focus on using machine learning and neural networks in question answering, reintroducing *semantic parsing* as a way of directly converting questions to formal queries or database queries. With respect to the kinds of things we were interested in in *cooperative question answering*, I think it's fallen to dialogue systems to pick up the slack on the kinds of interactions between a questioner and a respondent that question answering really requires.

On the other hand, I think we can still rely on the fact that problems that we tried to address and approaches that we tried to use 10 or 20 or 30 or even 40 years ago, and then abandoned for lack of machines powerful enough to support our efforts, are now getting rediscovered, with people make new progress. The optimist in me thinks, or at least hopes, that this will continue to be the case. So while question answering has changed, hopefully it will come back to where we started and go off with much more power behind it.

We should probably wrap up now because there's not that much time left. But before we do, I want to thank you, and Mark Johnson, and Anna Korhonen, and Barbara di Eugenio, and Candy Sidner, and Mark Steedman, all of whom helped me prepare for this interview. And finally, I would also like to express my ever-lasting gratitude to the first recipient of the Lifetime Achievement Award, Aravind Joshi, who, until his death in December 2017, was always a source of generosity and good humor and stimulating ideas and encouragement to all the people that he interacted with. We were all very lucky to have known him.

And thanks again for taking the time to interview me.

References

Agarwal, Shashank and Hong Yu. 2009. Automatically classifying sentences in full-text biomedical articles into introduction, methods, results and discussion. *Bioinformatics*, 25(23): 3174–3180.

Cheikes, Brant. 1991. *An Architecture for a Cooperative Respondent*. Ph.D. thesis, Department of Computer & Information Science, University of Pennsylvania.

Cheikes, Brant and Bonnie Webber. 1988. The design of a cooperative respondent. In *Proceedings of the Workshop on Architectures for Intelligent Interfaces*, pages 3–17, Monterey, CA.

Cohen, K. Bretonnel, Helen Johnson, Karin Verspoor, Christophe Roeder, and Lawrence Hunter. 2010. The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, 11.

Hirohata, Kenji, Naoki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. 2008. Identifying sections in scientific abstracts using conditional random fields. In

Proceedings of the 3rd International Joint Conference on Natural Language Processing, pages 381–388.

Joshi, Aravind, Bonnie Webber, and Ralph Weischedel. 1984. Preventing false inferences. In *Proceedings of COLING-84*, pages 134–138, Stanford, CA. DOI: `https://doi.org/10.3115/980431.980520`

Kaplan, Jerrold. 1982. Cooperative responses from a portable natural language database query system. In Brady, Michael and Robert Berwick, editors, *Computational Models of Discourse*, MIT Press, Cambridge MA, pages 167–208.

Kocijan, Vid, Oana-Maria Camburu, Ana-Maria Cretu, Yordan Yordanov, Phil Blunsom, and Thomas Lukasiewicz. 2019a. WikiCREM: A large unsupervised corpus for coreference resolution. In *Proceedings, Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4303–4312, Hong Kong. DOI: `https://doi.org/10.18653/v1/D19-1439`

Kocijan, Vid, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019b. A surprisingly robust trick for the Winograd schema challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4837–4842, Florence. DOI: `https://doi.org/10.18653/v1/P19-1478`

Liakata, Maria, Simone Teufel, Advaith Siddharthan, and Colin Batchelor. 2010. Corpora for the conceptualisation and zoning of scientific papers. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valletta, Malta. `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.724.8203&rep=rep1&type=pdf`

Lin, Jimmy, Damianos Karakos, Dina Demner-Fushman, and Sanjeev Khudanpur. 2006. Generative content models for structural analysis of medical abstracts. In *Proceedings of the HLT-NAACL Workshop on BioNLP*, pages 65–72. DOI: `https://doi.org/10.3115/1654415.1654427`

Lin, Ziheng, Hwee Tou Ng, and Min-Yen Kan. 2012. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering*, 20:151–184. DOI: `https://doi.org/10.1017/S1351324912000307`

Mays, Eric. 1984. *A Modal Temporal Logic for Reasoning about Changing Data Bases with Applications to Natural Language Question Answering*. Ph.D. thesis, Department of Computer and Information Science, University of Pennsylvania.

Mays, Eric, Aravind Joshi, and Bonnie Webber. 1982. Taking the initiative in natural language data base interactions: Monitoring as response. In *Proceedings of the European Conference on Artificial Intelligence*, pages 255–256, Orsay.

McKeown, Kathleen. 1985. *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Texts*. Cambridge University Press, Cambridge, England.

Mizuta, Yoko, Anna Korhonen, Tony Mullen, and Nigel Collier. 2006. Zone analysis in biology articles as a basis for information extraction. *International Journal of Medical Informatics*, 75:468–487. DOI: `https://doi.org/10.1016/j.ijmedinf.2005.06.013`, PMID: 16112609

Prasad, Rashmi, Bonnie Webber, and Aravind Joshi. 2014. Reflections on the Penn Discourse Treebank, comparable corpora and complementary annotation. *Computational Linguistics*, 40(4):921–950. DOI: `https://doi.org/10.1162/COLI_a_00204`

Rohde, Hannah, Anna Dickinson, Nathan Schneider, Christopher Clark, Annie Louis, and Bonnie Webber. 2017. Exploring substitutability through discourse adverbials and multiple judgments. In *Proceedings, 12th International Conference on Computational Semantics*, Montpellier.

Rohde, Hannah, Alexander Johnson, Nathan Schneider, and Bonnie Webber. 2018. Discourse coherence: Concurrent explicit and implicit relations. In *Proceedings of the 56th Annual Meeting of the ACL*. DOI: `https://doi.org/10.18653/v1/P18-1210`

Stede, Manfred. 2012. *Discourse Processing*. Morgan & Claypool Publishers.

Webber, Bonnie. 1979. *A Formal Approach to Discourse Anaphora*. Garland Press, New York. PhD thesis, Harvard.

Webber, Bonnie. 1982. So what can we talk about now? In Brady, Michael and Robert Berwick, editors, *Computational Models of Discourse*, MIT Press, Cambridge MA, pages 331–371.

Webber, Bonnie. 1986. Questions, answers and responses. In Brodie, Michael and John Mylopoulos, editors, *On Knowledge Base Systems*, Springer-Verlag, New York, pages 365–401. DOI: `https://doi.org/10.1007/978-1-4612-4980-1_30`

Webber, Bonnie, Markus Egg, and Valia Kordoni. 2012. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437–490. DOI: https://doi.org/10.1017/S1351324911000337

Webber, Bonnie, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The Penn Discourse Treebank 3.0 Annotation Manual, University of Pennsylvania. Available at https://catalog.ldc.upenn.edu/docs/LDC2019T05/PDTB3-Annotation-Manual.pdf.

Winograd, Terry. 1972. *Understanding Natural Language*. Academic Press, New York. Also published in *Cognitive Psychology*, 3:1(1972), pp. 1–191. DOI: https://doi.org/10.1016/0010-0285(72)90002-3

Woods, William. 1978. In Semantics and quantification in natural language question answering. *Advances in Computers*, 17. Academic Press, New York, pages 1–87. DOI: https://doi.org/10.1016/S0065-2458(08)60390-3

Woods, William, Ron Kaplan, and Bonnie Nash-Webber. 1972. The LUNAR sciences natural language information system: Final report, Bolt Beranek and Newman, Cambridge MA.

Xue, Nianwen, Hwee Tou Ng, Sameer Pradhan, Rashmi Prasad, Christopher Bryant, and Attapol Rutherford. 2015. The CoNLL-2015 shared task on shallow discourse parsing. In *Proceedings of the 19th Conference on Computational Natural Language Learning*, pages 1–16, Beijing. DOI: https://doi.org/10.18653/v1/K15-2001

Xue, Nianwen, Hwee Tou Ng, Sameer Pradhan, Attapol Rutherford, Bonnie Webber, Chuan Wang, and Hongmin Wang. 2016. CoNLL 2016 shared task on multilingual shallow discourse parsing. In *Proceedings of the 20th Conference on Computational Natural Language Learning*, pages 1–19, Berlin. DOI: https://doi.org/10.18653/v1/K16-2001