Circumscribing is not excluding: A response to Manning

Richard Crouch, Lauri Karttunen, Annie Zaenen

Palo Alto Research Center
Palo Alto, California 94305, USA

Manning [1] presents an extended critique of a paper by Zaenen, Karttunen and Crouch [2] (henceforth ZKC), which was a commentary on the PASCAL Recognizing Textual Entailment (RTE) challenge [3]. The ZKC paper was entitled "Local textual inference: can it be defined or circumscribed?" and argued that the PASCAL data should have been better defined and circumscribed so as to distinguish between different forms of textual inference. Some misunderstanding has arisen, and Manning takes us to be in favor of excluding certain forms of textual inference rather than just distinguishing between them. In this note we clarify how the considerations put forward in ZKC serve not to narrow the textual inference task, as Manning claims, but to broaden it. We illustrate this by discussion of the AQUAINT Knowledge-based Evaluation (KBEval) data and annotation guidelines [4][1], which we played a large role in formulating.

Manning's principal complaint is that we seek to narrow the definition of textual inference "so as to exclude many of the inferences that humans make and many of the inferences that are needed for operational use of robust textual inference" ([1] p. 2). In fact the opposite is true. In ZKC we argued only that the impact of world knowledge and plausible inference should have been *circumscribed* in the PASCAL RTE data, not that it should have been *excluded*. That is, the data annotation would have benefited from drawing lines around and distinguishing between inferences based solely on linguistic knowledge and those also based on world knowledge, and between strictly valid inferences and merely plausible inferences. It was not our intent to exclude all but one form of inference, and this should be evident from the AQUAINT KBEval annotation guidelines. A comparison of the PASCAL and AQUAINT annotations schemes moreover shows that the distinctions ZKC argued for supports a broad and open definition of textual inference.

Recall that the PASCAL data provides *Text/Hypothesis* pairs, annotated with whether the hypothesis does, or does not, follow from the text. The data is also divided into seven classes (IR, CD, RC, QA, IE, MT, PP) reflecting the kind of application the data is relevant to.

[1] ZKC only offers two paragraphs of recommendations on improving the PASCAL data annotation. However, the annotation guidelines for the AQUAINT KBEval are spelled out in detail in [4].

The KBEval annotation scheme is a more refined version of the PASCAL scheme, applicable to all the PASCAL RTE data.[2] Rather than just stating whether the hypothesis does or does not follow from the text, the pairs are annotated along the following dimensions

- **Polarity**: *true*, *false* or *unknown*
  If true, the hypothesis follows from the text. If false, the hypothesis is contradicted by the text. If unknown, the hypothesis neither follows from nor is contradicted by the text.
- **Force**: *strict* or *plausible*
  For pairs with polarity true or false, could additional information or passages consistent with the text lead to a change in the polarity? If not, the force is strict. Otherwise it is plausible.
- **Source**: *linguistic* or *world*
  Would any competent speaker of the language, no matter how ignorant otherwise, make the same judgment about the polarity. If not, the source of the inference lies in world knowledge. Otherwise, it is based solely on knowledge of the language.

Polarity is a three-way distinction refining PASCAL's *follows/does not follow* two-way distinction. Force and Source are classification dimensions not present in PASCAL. There are also some further optional annotations, e.g. a textual description of any world knowledge assumed[3], a description of the context under which the inference holds, and further assumptions about interpretations of ambiguities. We will not discuss these here.

Contrary to Manning's assertions, it should be apparent from this classification scheme that the KBEval annotations are aiming to do **none** of the following:

1. Exclude inferences based on world knowledge ([1] pp. 3–5). As the annotation scheme makes clear, the aim is to distinguish inferences based on world knowledge from other inferences, not to exclude them[4]. Much of the discussion in [1] concerns how the PASCAL RTE challenge would have been weakened if all inferences based on world knowledge had been removed. This is true but rather beside the point, since no one was arguing for their exclusion. Manning seems to agree with the proponents of RTE in that the world knowledge should be *common* knowledge and understanding. ZKC are not the only ones that find the first RTE set is a bit broad in its view of what constitutes common knowledge, and could do with circumscribing it better: Newman et al. 2005 [5] and Bayer et al. 2005 [6] make a similar point.
2. Exclude particularized conversational implicatures, whereby world knowledge can be used to infer a speaker's intended meaning for a piece of text

[2] The KBEval annotations are also modified to apply to Passage/Question/Answer triples, but to ease comparision we will here apply it to PASCAL's Text/Hypothesis pairs.

[3] Thus providing the additional background context argued for in Manning's proposed modification to PASCAL (p. 11).

[4] How easy it is to make this circumscription is something we return to below.

([1] p. 7). These examples of plausible inference based on world knowledge are not ruled out. ZKC merely pointed out that without a sufficiently circumscribed statement of background context these implicatures are hard to calculate reliably, and so are problematic given the PASCAL annotation scheme.

3. Exclude conclusions drawn from statements made in reported speech or modal contexts ([1] pp. 5–6). This is similar to particularized conversational implicatures. In a recent related paper [7], two of us discuss the various ways authors can signal their commitment to reported statements and we state explicitly that this knowledge should be combined with knowledge about the trustworthiness of the source. We do not propose to ignore the trustworthiness of the source. Rather, this should be circumscribed as a factor that contributes to calculating the trustworthiness of the statement.
4. Commit to an unnatural, narrow, academic notion of inference that does not correspond to everyday use ([1] p. 2). A variety of different types of inference (strict/plausible and linguistic/world) are distinguished, in a way that is certainly no less intuitive than Manning's appeal to the inferences that would drawn by a person of common knowledge and understanding. However, Manning's appeal to inferences drawn by the "common man" is both suspect and unduly restrictive[5]. Common human reasoning exhibits some recurring patterns of fallacious inference. Actually occurring human reasoning constitutes a legitimate and subfield of psychology and perhaps of textual inference. But one would not want a science of textual inference to entrench these fallacies as somehow being "empirically correct". This is especially relevant for applications to do with intelligence analysis, where it is a major effort to train analysts so that they don't make the same kinds of mistakes that the common man makes.
5. Commit to a particular linguistic theory of syntax, semantics or pragmatics ([1] p. 7). The annotation scheme makes no reference to any terms of art from such theories (e.g. implicature, presupposition, downward monotonicity). The use of such terms in ZKC was a convenient way of referring to naturally occurring classes of data that were missing or badly under-represented in the PASCAL data. The particular way of distinguishing these classes may have been theoretically loaded, but the absence of the data is not. Bayer et al. [6], for example, note that 94% of the positive entailments in the RTE dev2 set were mere paraphrases as opposed to classic entailments.

The claim that the authors of ZKC intend to do all of (1)–(5) above seems to be based on a fundamental misconstrual of what the ZKC paper and the KBEval annotation scheme is about. How might this misunderstanding have arisen?

It is important to distinguish between an annotation scheme, and the way that the scheme can be used to support evaluations. From a machine learning

[5] One might question the assertion that the PASCAL data represent the judgments of the common man. They represent the judgments of graduate students. A serious study of the judgments of the common man would throw valuable light on why statements lead so often to unwarranted conclusions.

perspective, not everything that is annotated needs to be learned. For example, it was presumably not intended that PASCAL's seven-fold IR–PP classification should be one of the outputs of any ML system. Rather this annotation was intended either to allow an informative breakdown and analysis of results, or to allow a particular system to target a certain subset of the data.

The same holds of the KBEval annotations. The annotation guidelines are explicit that the only mandatory system output is a judgment of polarity (true, false or unknown). It leaves open the question of whether judgments of force and source could optionally be system outputs, or whether, as in the AQUAINT pilot evaluation, these annotations just provide the basis for a more informative breakdown and analysis of results.

The KBEval annotation scheme was designed to allow a single data set to support a number of different forms of evaluation, applicable to a range of systems built on different principles. For example, component systems that assume a linguistic/world knowledge distinction and only deal with strict linguistic inferences would be expected to perform significantly better on the strict linguistic portion of the data whereas systems with a more monolithic conception of inference could undergo an evaluation that simply ignores the force and source annotations in the data set.

Perhaps it is the ability of the KBEval annotations to support different kinds of evaluation that has led to misunderstanding. To see how this might arise, imagine combining the following two assumptions, the first being perfectly reasonable but the second being misplaced. The first assumption is that there is a monolithic conception of textual inference; distinctions between world v. linguistic knowledge and strict v. plausible inference cannot coherently be drawn, and are epiphenomena of semanticists' devising.[6] The second assumption is that any system has to be able to produce as output all the classifications present in the annotation. The second assumption, combined with the classification dimensions in the KBEval data, would then force the monolithic system to make precisely the classifications that the first assumption denies are possible. It might thus seem that the KBEval data is some kind of formal semanticists' conspiracy to exclude certain approaches to modeling textual inference. Of course, it is no such thing, and the rational response is not to junk either the monolithic conception of textual inference or the KBEval data. It is to junk the faulty assumption that all annotation dimensions have to be directly reflected in system output.

The case of the PASCAL RTE annotation scheme provides an interesting contrast. While a system does not have to output all dimensions in an annotation scheme, it should surely output at least one of them. Leaving aside PASCAL's IR–PP task classification, the only remaining classification is that of polarity. So from the point of view of a component system focused on strict linguistic inference, any PASCAL-based evaluation would force it to embrace exactly the monolithic conception of textual inference that such a system denies.

We agree with Manning when he states that (p. 5)

[6] This is not an assumption to which we subscribe, but it is a perfectly coherent one to make.

> I think it is important to develop a common playing field where linguistic processing technologies and KR&R technologies can be fruitfully combined, and the values of different components can be carefully measured.

The disagreement is with his implication that ZKC and the KBEval annotation scheme gives a “narrow definition that serves to undermine rather than encourage . . . a new rapprochement between the human language technology and knowledge representation and reasoning communities.” On the contrary, ZKC was an attempt to explicate what a broad approach to evaluation would be and the KBEval annotations illustrate that view: they are compatible with a monolithic account of inference, as well other alternatives. They provide a basis for evaluating the quality of different components both on their own terms, and in terms of their overall contribution to the task of recognizing the full range of textual inferences. The PASCAL RTE data as it stands is a valuable step towards a rapprochement, and further articulation will help bring in a number of players with different approaches and goals that might be otherwise excluded.

Appendix 1: World v. Linguistic Knowledge

There are a number of other points raised by Manning that deserve some kind of response. The most obvious regards the vexed question of world v. linguistic knowledge. It is certainly true that ZKC “try to cleave off a linguistic textual inference task that excludes common sense and basic world knowledge,” ([1] p. 4) and it may eventually turn out that this is “precisely the wrong thing to do from the perspective of developing the necessary science for text understanding.” The point is that you won’t find out that it’s the wrong thing to do unless you try. There is nothing in ZKC or the KBEval annotations that penalizes other researchers who do not make the cleavage between language and the world; so there seems little harm in letting ZKC go their own way. The problem with the PASCAL data is that it provides no room on the common playground allowing ZKC to try their experiment, and so does penalize them.

Some might hold that the ZKC cleavage between language and the world is doomed to an early failure because the distinction cannot reliably be made. As Manning states ([1] p. 6) “there are always going to be borderline cases in a classification task,” and this is perhaps nowhere more true than in distinguishing linguistic from world knowledge. But difficult borderline cases are not evidence of a fictitious border; unless, that is, the borderline cases are numerous enough to occupy most of the territory. It seems hard to deny that there are clear-cut examples of purely linguistic inference

> *Text:* Indonesia is the largest archipelagic nation in the world, consisting of 13,670 islands.
> *Hypothesis:* 13,670 islands make up Indonesia.

and that there are examples requiring extensive amounts of world knowledge

> *Text:* The country’s largest employer, Wal-Mart Stores Inc., is being sued by a number of its female employees who claim they were kept out of

jobs in management because they are women.
Hypothesis: Wal-Mart sued for sexual discrimination.

There are also borderline cases

Text: The domestic cat is a mammal of the genus *felix.*
Hypothesis: Domestic cats suckle their young.

(Is it a fact about the world that mammals suckle their young, or is it part of the definition of "mammal"?). The question is, how extensive is the borderline region?

This is a question that we can begin to answer on the basis of inter-annotator agreement on the KBEval data. Manning [1] cites agreement on 74 out of 76 examples for the PARC contribution to the KBEval dataset (97%). However, this is not a reliable indicator given that these were relatively simple examples, all of which were claimed to illustrate purely linguistic inference. Anecdotally, it appears that the linguistic / world knowledge distinction did not provoke the disagreement one might have expected in the other contributions to the KBEval dataset, but this should be subjected to formal scrutiny.

Appendix 2: Laboratory Examples

Discussion of the PARC contribution to the KBEval dataset raises the issue of the role of hand constructed, laboratory examples. There are two things to be said in response to Manning's paper.

First, it is a mistake to see the PARC KBEval contribution as evidence that ZKC wished to restrict all evaluation to strict linguistic inferences([1] p. 10). The PARC contribution did indeed consist entirely of hand constructed examples targeting linguistic inferences that were predominantly strict rather than plausible. Taken on its own, this would provide a narrowly limited range of examples. But the PARC data was not intended to be taken on its own. A variety of other sites provided further annotated data, covering a wide range of strict and plausible, and linguistic and world-based inferences. Far from being more restricted than PASCAL, it would have been quite possible to include all the (reannotated) PASCAL data in the KBEval data, had this been in the form of question-answer pairs.

Second, there is a point of genuine dispute between us and Manning over whether hand constructed examples are legitimate in inference evaluation data. Manning argues for restricting the data solely to naturally occurring examples (although he reluctantly concedes that most hypothesis statements in the PASCAL data were in fact hand constructed). We argue for a broader approach that legitimizes both naturally occurring (or field) examples, and hand constructed (or laboratory) examples. Most scientific disciplines acknowledge the utility of combining laboratory and field studies: Laboratory controls make it easier to isolate and identify core phenomena, but the insights gained from controlled experimentation must then be shown to have field validity. The PARC portion of the KBEval dataset is reported to have helped some in debugging and advancing their technology [8], underscoring the practical and scientific utility of allowing both field and laboratory data.

Appendix 3: Manning on the evolution of semantics and pragmatics

Apart from misunderstandings about our opinions and aims, Manning [1] states that several of the distinctions that ZKC refer to are no longer generally accepted by semanticists and pragmaticists. Since ZKC was not aiming to impose the theories and distinctions that its authors have elsewhere promoted, Manning's claims are of marginal relevance to the present discussion. However, on the grounds that several decades of work on semantics and pragmatics might just have something to contribute to our understanding of textual inference, it seems worth clearing up some of his misconceptions about this area.

It is of course true that semantics and pragmatics, like all sciences, have their unsolved problems and internal debates but Manning misconstrues some of the controversies that he refers to and overestimates the importance they have for the points made in ZKC. Referring to the taxonomy of inference types dating back to the "Golden Age of Pure Pragmatics" (Horn [9]), Manning comments "Since that time, many problems for the accounts proposed then have accumulated, and as a result, today there is no longer a broad consensus of opinion supporting the above taxonomy (even among Anglo-American philosophers)." (p. 7) This is a misrepresentation. The citations of Horn [9] and Bach [10] do not support such a claim. Manning implies that Horn has changed his mind about scalar implicatures, he hasn't. Horn has modified his earlier analysis with respect to cardinals: "These observations suggest the need for a mixed approach on which cardinal values demand an enriched-content analysis while other scalar predications continue to submit to a standard neo-Gricean treatment on which they are lower-bounded by their literal content and upper-bounded, in default contexts, by implicature". [9] p. 4.

Manning suggests that Bach's opinion invalidates our discussion of conventional implicatures but what Bach does is classify conventional implicatures in a different way, packaging them as part of 'what is said', in contrast with the treatment given in Karttunen&Peters 1979 [11] and Potts 2005 [12]. As for the general issue Bach [13] concludes "The semantic-pragmatic distinction is a well-defined and theoretically warranted distinction" (p. 42). More generally Manning notes that not everybody draws the distinction between pragmatics and semantics in the way we seem to imply. That debate is, however, irrelevant to the practical relevancy of the distinctions that we draw[7].

[7] In a footnote (p. 7) Manning suggests, quoting Recanati [14], that according to the traditional account, the literal meaning of "You are not going to die" is that the hearer is immortal. This is a gross mischaracterization of traditional accounts of tense and aspect. Only the most sketchy and introductory discussions might omit to point out that such statements have to be relativized to a context.

References

1. Manning, C.D.: Local textual inference: it's hard to circumscribe, but you know it when you see it — and NLP needs it. (2006) Unpublished manuscript. February 25. http://nlp.stanford.edu/~manning/papers/LocalTextualInference.pdf.
2. Zaenen, A., Karttunen, L., Crouch, R.: Local textual inference: can it be defined or circumscribed? In: ACL 2005 Workshop on Empirical Modelling of Semantic Equivalence and Entailment, Ann Arbor, Michigan (2005) http://www2.parc.com/istl/members/karttune/publications/acl2005workshop.pdf.
3. Dagan, I., Glickman, O., Magnini, B.: The PASCAL recognising textual entailment challenge. In: Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment. (2005) http://www.cs.biu.ac.il/~glikmao/rte05/dagan_et_al.pdf.
4. Crouch, R., Sauri, R., Fowler, A.: AQUAINT pilot knowledge-based evaluation: Annotation guidelines. (2005) Manuscript. http://www2.parc.com/istl/groups/nltt/papers/aquaint_kb_pilot_evaluation_guide.pdf.
5. Newman, E., Stokes, N., Dunnion, J., Carthy, J.: UCD IIRG approach to the textual entailment challenge. In: Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment. (2005) http://www.cs.biu.ac.il/~glikmao/rte05/newman_et_al.pdf.
6. Bayer, S., Burger, J., Ferro, L., Henderson, J., Yeh, A.: Mitre's submissions to the EU pascal RTE challenge. In: Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment. (2005) http://www.cs.biu.ac.il/~glikmao/rte05/bayer_et_al.pdf.
7. Karttunen, L., Zaenen, A.: Veridicity. In G. Katz, J. Pustejovsky, F.S., ed.: Annotating, Extracting and Reasoning about Time and Events. Volume 05151 of Dagstuhl Seminar Proceedings., Dagstuhl, Germany (2005) http://www2.parc.com/istl/members/karttune/publications/Veridicity.pdf.
8. de Salvo Braz, R., Girju, R., Punyakanok, V., Roth, D., Sammons, M.: An inference model for semantic entailment in natural language. In: Proceedings of the Workshop on Knowledge and Reasoning for Answering Questions, IJCAI–05. (2005) http://l2r.cs.uiuc.edu/~danr/Papers/BGPRS05.pdf.
9. Horn, L.R.: The border wars: a neo-Gricean perspective. In Turner, K., von Heusinger, K., eds.: Where Semantics Meets Pragmatics. Elsevier, Amsterdam, the Netherlands (2006) http://www.yale.edu/linguist/faculty/doc/horn_border.doc.
10. Bach, K.: The myth of conventional implicature. Linguistics and Philosophy **22** (1999) 367–421 http://userwww.sfsu.edu/~kbach/Myth.htm.
11. Karttunen, L., Peters, S.: Conventional implicature. In Oh, C.K., Dinneen, D.A., eds.: Syntax and Semantics, Volume 11: Presupposition. Academic Press, New York (1979) 1–56 http://www2.parc.com/istl/members/karttune/publications/ConvImp.pdf.
12. Potts, C.: The Logic of Conventional Implicatures. Cambrige University Press, Cambridge, United Kingdom (2005)
13. Bach, K.: Minding the gap. In Bianchi, C., ed.: The Semantics Pragmatics Distinction. CSLI Publications, Stanford University. Palo Alto, California (2004) 27–43 http://userwww.sfsu.edu/~kbach/MindingtheGap.pdf.
14. Recanati, F.: Literal Meaning. Cambridge University Press, Cambridge, United Kingdom (2004)