LOCAL TEXTUAL INFERENCE: IT'S HARD TO CIRCUMSCRIBE, BUT YOU KNOW IT WHEN YOU SEE IT – AND NLP NEEDS IT

CHRISTOPHER D. MANNING

STANFORD UNIVERSITY

25 FEBRUARY 2006

Technology for *local textual inference* is central to producing a next generation of intelligent yet robust human language processing systems. One can think of it as Information Retrieval++. It is needed for a search on *male fertility may be affected by use of cell phones* to match a document saying *Startling new research into mobile phones suggests they can reduce a man's sperm count up to 30%*, despite the fact that the only word overlap is *phones*. But textual inference is useful more broadly. It is an enabling technology for applications of document interpretation, such as customer response management, where one would like to conclude from the message *My Squeezebox regularly skips during music playback* that *Sender has set up Squeezebox* and *Sender can hear music through Squeezebox*, and information extraction, where from the text *Jorma Ollila joined Nokia in 1985 and held a variety of key management positions before taking the helm in 1992*, one wants to extract that *Jorma Ollila has served as the CEO of Nokia*, a relation that might be more formally denoted as *role(CEO, Nokia, Jorma Ollila)*. Textual inference is a difficult problem (as the results from early evaluations have shown): current systems do statistically better than random guessing, but not by very much. Nevertheless, it is also an area where there is promising developing technology and a good deal of natural language community interest. In other words, it is an ideal research problem. To further this research agenda, data sets have been constructed to assess textual inference systems. This paper examines how the task of textual inference has been and should be defined and discusses what kind of evaluation data is appropriate for the task.[1]

DEFINING THE TASK OF TEXTUAL INFERENCE

The Pascal RTE1 Challenge

In 2005, the First PASCAL Recognizing Textual Entailment Challenge (RTE1)[2] sought to test whether computer systems can draw appropriate inferences from a short piece of text, as humans can, where the hypothesis tested is also expressed in textual form. For particular concrete applications, the hypothesis might be analogous to a query to an information retrieval system or a statement expressing a putative answer to a question in a question-answering system. For example:

> *Text:* Soprano's Square: Milan, Italy, home of the famed La Scala opera house, honored soprano Maria Callas on Wednesday when it renamed a new square after the diva.
>
> *Hypothesis:* La Scala opera house is located in Milan, Italy. TRUE. (RTE1 ID: 565)

The text for an item is a short passage, usually just one sentence. The hypothesis is a single sentence. There were only a couple of technical provisos for the task. The text is assumed to be from a

[1] My thanks to Andrew Ng, for emphasizing the importance of using sensory data in AI, and to Stanley Peters, for a useful discussion on literal meaning versus speaker meaning.

[2] Details can be found at *http://www.pascal-network.org/Challenges/RTE/*.

trustworthy source. Tense is to be ignored, to allow matching over texts written at different times.[3] Finally, one is to assume that compatible referring expressions in the text and hypothesis have the same sense and reference in the absence of evidence to the contrary. That is, if the text and hypothesis both mention "Paris" one should not allow that one is talking about France and the other about Paris, Texas, with the upshot that the text has no bearing on the truth of the hypothesis.

Two opposing perspectives

A recent paper by Zaenen et al. (2005) – henceforth ZKC – attempts to circumscribe what phenomena should appear in a test of local textual inference for advanced human language understanding systems. They make their proposal "in the hope of getting a discussion going." This paper takes the bait. Both their and my discussion is mainly in the context of the PASCAL RTE1 challenge. However, like ZKC, I prefer the name *local textual inference* to *recognizing textual entailment*, for reasons discussed below. I will also refer to the Knowledge-based Inference Pilot organized within the US Government ARDA AQUAINT program, which both myself and ZKC took part in.

ZKC are right to point out the importance of certain logical/semantic distinctions to robust textual inference – distinctions that many working in the area were initially insufficiently aware of. For instance, they are right to emphasize how many methods used in RTE1 only treat upward monotonic entailments (a point also made in our paper, Haghighi et al. (2005: 392)). Their paper gives a useful rendition and taxonomy of the standard formal semantic treatment of drawing inferences from text. Nevertheless, I submit that ZKC improperly seek to narrowly circumscribe the task of local 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. Moreover, their narrow definition serves to undermine rather than encourage the possibilities for a new rapprochement between the human language technology and knowledge representation and reasoning communities.

I believe that the right way to design a textual inference task is to adopt as the standard of inference what a human would be happy to infer from a piece of text. In particular, items would be assessed by people that are awake, careful, moderately intelligent and informed, and with reasonable document interpretation expertise, but not by semanticists or similar academics. These people would use whatever background and real world knowledge that they usually bring to interpreting texts. This is the kind of pool and procedure that the NIST TREC evaluation has always used for determining the relevance of results returned by information retrieval and question answering (QA) systems. The texts for the task should be short passages of naturally occurring text. I feel it is vital to keep the task grounded in real data. But I think the hypothesis should not always be authentic text. The hypothesis is a probe, and one sometimes wants to probe whether a system understands particular things. It would often be difficult or impossible to find authentic text that probed these things. Nevertheless, to improve task realism and grounding, it is desirable for the hypothesis to be drawn from motivating tasks and constructed independently of the text as much of the time as is practical.

The last paragraph mainly sounds to be arguing for good experimental practice. Good experimental practice is very desirable, but beyond this, such a design is the best procedure for reaching the two key goals of (i) a robust textual inference task that is reflective of and sensitive to plausible operational system needs, and (ii) providing a new venue for interaction between Natural Language Processing (NLP) and Knowledge Representation and Reasoning (KR&R). Within this framework, NLP people can do robust language processing to get things into a form that KR&R people can use, while KR&R people can show the value of using knowledge bases and reasoning that go beyond the shallow bottom-up semantics of most current NLP systems.[4] Local textual inference is a clear task, with a natural and straightforward, human-understandable evaluation procedure. The task avoids a commitment to any particular knowledge representation, but it allows people to exploit any knowledge representation and reasoning mechanisms that help for the task at hand.

[3] This is a limitation, since one would also like to be able to assess temporal inference. But it was done for practical reasons in this initial dataset, because collected texts often used different tenses for events.

[4] The best representative NLP systems are the QA systems of (Harabagiu et al. 2000, Moldovan et al. 2003).

Was world knowledge part of the Pascal RTE task?

ZKC open (p. 31) by suggesting that "The PASCAL initiative on 'textual entailment' had the excellent idea of proposing a competition testing NLP systems on their ability to understand language separate from the ability to cope with world knowledge." I think this is a fundamental misconstrual of what PASCAL RTE aims to do, reflecting rather what ZKC wish it were doing, and that a lot of their unhappiness with PASCAL follows from this misconstrual. Below I present my understanding of what is being tested, and argue that the organizers made roughly the right decision by designing things the way they did. I was not one of the organizers, but I will provide some textual evidence to support the contention that the organizers had in mind a definition of inference much like the one I argue for, whereas ZKC provide no source to support their characterization.

A clear statement that the task is not separated from the use of world knowledge appears in the summary paper of the organizers (Dagan et al. 2005: 1): "We say that T entails H if the meaning of H can be inferred from the meaning of T, as would typically be interpreted by people. This somewhat informal definition is based on (and assumes) common human understanding of language as well as common background knowledge." Almost identical text also appeared in the instructions for the challenge.[5] This position has been amplified in the second round RTE instructions: "Our definition of entailment allows presupposition of common knowledge, such as: a company has a CEO, a CEO is an employee of the company, an employee is a person, etc. For instance, in example #6, the entailment depends on knowing that the president of a country is also a citizen of that country."[6] The basis all along seems to have been that world knowledge *is* part of the task.[7]

However, the allowed use of world knowledge is a background use. The text must establish the (likely) truth of the major proposition of the hypothesis for it to be true. The notion of inference is not material implication (where a true hypothesis is implied by any text (whether true or false), but somewhat more like relevance logic (cf. Restall & Dunn 2002). Hence rather than viewing a text/hypothesis pair as "true" or "false", one might much prefer to say that the hypothesis "follows" or "does not follow" from the text, and I will use these terms below. Thus an example like the following is judged false ("does not follow"), even though Grozny is the capital of Chechnya:

> *Text:* While civilians ran for cover or fled to the countryside, Russian forces were seen edging their artillery guns closer to Grozny, and Chechen fighters were offering little resistance.
>
> *Hypothesis:* Grozny is the capital of Chechnya. FALSE. (RTE1 ID: 583)

One way of thinking about whether an hypothesis follows from a text is: if a person asked you for a piece of text that establishes a certain hypothesis, then would showing them the given text satisfy the person. From an operational perspective, this seems just what we want. For instance, think of applications like passage retrieval, question answering, or event extraction. It is not useful to provide any document at all as a justification for something that we know to be true from a database table or knowledge base. On the other hand, it is very reasonable to return the following text in support of the given hypothesis, even though its sufficiency is dependent on knowledge that Paris is in France:

> *Text:* The Mona Lisa, painted by Leonardo da Vinci from 1503-1506, hangs in Paris' Louvre Museum.
>
> *Hypothesis:* The Mona Lisa is in France. FOLLOWS. (RTE1 ID: 153)

Another important aspect of the setting is that an hypothesis is taken to follow from a text when the hypothesis is highly plausible given the text, even if it is not a logical entailment of the text (and any needed background knowledge). Here is an example:

> *Text:* The anti-terrorist court found two men guilty of murdering Shapour Bakhtiar and his secretary Sorush Katibeh, who were found with their throats cut in August 1991.

---

[5] See *http://www.pascal-network.org/Challenges/RTE/Instructions/*

[6] See *http://www.pascal-network.org/Challenges/RTE2/Instructions/*

[7] These statements are backed up by the annotated data. Many items require background knowledge, for example, RTE1 ID: 153, cited below, requires knowing that Paris and/or the Louvre Museum is in France.

Hypothesis: Shapour Bakhtiar died in 1991. FOLLOWS. (RTE1 ID: 579)

It is not an entailment that Shapour Bakhtiar died in 1991: he could have been killed in 1990 or even earlier, and it just took a very long time for anyone to find the bodies. But this is *extremely* unlikely given our understanding of how the world works. In talks, Ido Dagan, one of the PASCAL RTE organizers, has stressed that one wants to include "almost certain" conclusions, and that this is likely to be important for applications. I agree: in real texts, it is just very often the case that something almost certainly follows but the inference requires additional assumptions that are highly plausible but not necessary. But this is the clear reason to prefer the name *textual inference* to *entailment:* entailment is a technical term in logic, which means that a conclusion must necessarily follow from premises in *every possible* situation in which the premises are true. Within the RTE framework, the set of possible premises is not circumscribed but we are at any rate allowed to go beyond them to conclude things that are very reasonable but not strictly necessary.

Is it reasonable to include world knowledge (without carefully circumscribing it)?

ZKC present rather ambivalent views about the role of world knowledge in a textual inference task, initially saying that they do not want any, but gradually admitting that it is impossible to filter out. They start their paper with a strong statement that world knowledge is out of bounds (p.31): "This [their interpretation of the PASCAL initiative – CDM] is obviously a welcome endeavor: NLP systems cannot be held responsible for knowledge of what goes on in the world but no NLP system can claim to 'understand' language if it can't cope with textual inferences." However, later in their paper, they seem to weaken or even concede this point. On p. 33, particular delimited kinds of world knowledge are admitted: "Whether [temporal and spatial knowledge] is linguistic knowledge or world knowledge might not be totally clear but it is clear that one wants this information to be part of what textual entailment can draw upon." On p. 34, they go further, writing: "But even in a task that tries to separate out linguistic knowledge from world knowledge, it is not possible to avoid the latter completely. There is world knowledge that underlies just about everything we say or write." They essentially retreat to suggesting that the problem is not the use of world knowledge but that the boundaries of the allowed world knowledge are not defined (p. 34): "Then there is knowledge that is commonly available and static, e.g. that Baghdad is in Iraq. It seems pointless to us to exclude the appeal to such knowledge from the test suite but it would be good to define it more explicitly."

Including world knowledge has a practical basis, as ZKC reluctantly conclude: most people believe there is little hope of separating linguistic versus world knowledge. For the lexical case, this issue has been discussed under the rubric of "dictionaries vs. encyclopedias". Among others, Eco (1984) argues that dictionaries cannot successfully be distinguished from encyclopedias, writing that the dictionary dissolves "into an unordered and unrestricted galaxy of pieces of world knowledge", and Wierzbicka (1995) adopts an expansive view of dictionary definitions where much cultural and world knowledge is encoded within them. I will additionally argue that trying to cleave off a linguistic textual inference task that excludes common sense and basic world knowledge is precisely the wrong thing to do from the perspective of developing the necessary science for text understanding.

Would the task be improved by providing needed background world knowledge (in some textual or formalized form)? Not precisely delineating the admissible world knowledge in the PASCAL RTE task was doubtless partly a practical decision: as anyone who has followed the progress of the Cyc project (Lenat and Guha 1990) knows, the amount of commonsense and general world knowledge is vast and not easily delineated. It is much easier to stick with saying world knowledge is things that most people know. But, beyond practicality, this decision sets up an interesting task structure, conducive to important future research in several fields.

Within PASCAL RTE, the general conception is that any precise, event or domain specific facts needed to assess the hypothesis must be present in the text. This leaves to world knowledge precisely the role that the Cyc project originally had – providing all the necessary background assumptions and linkages to allow inferences to go through as they would for human beings – whereas in practice the Cyc project has often deviated from this goal and proceeded to build up large knowledge bases for

particular application-specific needs. Rejuvenating this original goal of enabling common sense inference is a promising direction for KR&R and a good basis for linking modern NLP with KR&R. It also leaves open knowledge acquisition from text as a useful and important research problem. If all the needed facts were provided next to a text, interesting natural language understanding problems might remain, but the knowledge acquisition and reasoning problems would be trivialized.

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 value of different components can be carefully measured. In particular, I would promote evaluating KR&R in a context that is still directly connected to raw sensory inputs, rather than a context where it works on knowledge hand-encoded by humans. Artificial Intelligence has always gone astray when it has insulated itself from using real sensory data as the input to systems. The PASCAL RTE task is a good candidate task for this: many of the problems require reasoning about the world, and I think it is fair to say that the only questions that current systems get right (other than as a lucky guess) are those for which no significant reasoning and information combination is required in getting from the text to the hypothesis. Thus it is a task where NLP needs help from KR&R.

How real-world textual inference differs from standard logical semantics

Two examples: modals and reported speech

Standard theories of linguistic semantics are ill-suited in many cases to modeling the inferences that people draw from texts (the task at which PASCAL RTE is aimed). Modal verbs provide one example. Use of *may* or *can* is logically taken to indicate only that something is a possible state of affairs (that it is true in some possible world); nothing can be concluded about whether the state of affairs holds in the real world. But, in practice, people often write text with *may* or *can* expecting the reader to take the clauses as true. One sees this interpretation at work in several of the PASCAL pairs:

> *Text:* Researchers at the Harvard School of Public Health say that people who drink coffee may be doing a lot more than keeping themselves awake - this kind of consumption apparently also can help reduce the risk of diseases.
>
> *Hypothesis:* Coffee drinking has health benefits. FOLLOWS. (RTE1 ID: 19)
>
> *Text:* Eating lots of foods that are a good source of fiber may keep your blood glucose from rising too fast after you eat.
>
> *Hypothesis:* Fiber improves blood sugar control. FOLLOWS. (RTE1 ID: 20)

I take the FOLLOWS judgements of the human annotators as being correct here. The presence of *may* or *can* is merely a form of hedging – a way of leaving open the possibility that further contradictory evidence might yet emerge. Researchers, just like politicians, like to hedge.

Another example concerns speech act verbs. In talks, Ido Dagan differentiates strict logical entailment from real-world inference and mentions how "Strict entailment … doesn't account for some uncertainty allowed in applications." He gives as an example a non-factive report, which he describes as a TRUE inference:

> *Text:* According to the Encyclopedia Britannica, Indonesia is the largest archipelagic nation in the world, consisting of 13,670 islands.
>
> *Hypothesis:* 13,670 islands make up Indonesia. FOLLOWS. (RTE1 ID: 605)

Contrast the viewpoint of ZKC presenting an exactly parallel *According to* construction, where they argue that the truth of the main clause does not follow (p. 33):

> "It is important to point out that the syntactic structure doesn't guide the interpretation here. Consider the following contrast:
>
> (12) As the press reported, Ames was a successful spy.
>
> conventionally implicates that Ames was a successful spy, but
>
> (13) According to the press, Ames was a successful spy.

does not."

It is the nature of the PASCAL data that we are to assume that the statement made in the text should be accepted as true (as from a trustworthy author, as ZKC more finely put it on p. 31). However, as soon as there is an embedded verb of report, such as:

> The American State Department announced that Russia recalled her ambassador to the United States 'for consultation' due to the bombing operations on Iraq. (RTE1 ID: 352)

then the standard linguistic account assumed by ZKC says that we can conclude nothing about Russia's actions. Someone reported something; it may or may not be true. But this just isn't a useful approach in the real world. If we ascribe no truth to anything that we learn of by reports, then we would gain very little knowledge ever. On the other hand, we cannot be naïve and believe everything. The process required is just what every informed reader does (and one which I presume intelligence analysts pay a lot of attention to): they carefully evaluate the source with respect to the report and decide whether they think that the report should be believed. In the absence of contradictory evidence, most people accept most reports from major news outlets and government agencies.

An example that particularly rankles ZWC is:

> *Text:* A statement said to be from al Qaida, claimed the terror group had killed one American and kidnapped another in Riyadh.
>
> *Hypothesis:* A U.S. citizen working in Riyadh has been kidnapped after a colleague in the same company was shot dead yesterday. FOLLOWS. (RTE1 ID: 775)

This example is deemed an error by ZKC: an assumption that the author of the text is trustworthy should not be extended to cited sources, especially when the verb used is *claimed*. I admit that this example is not completely uncontroversial[8] – there are always going to be borderline cases in a classification task – but it actually seems to me not too unreasonable in the context of real world interpretation. It just happens to be the case that Al Qaeda is not prone to making announcements and claims that are not factual. Indeed, my understanding is that people use the absence of a claim of responsibility from Al Qaeda as good evidence that Al Qaeda was not involved in an operation. The reasoning here clearly involves extensive world knowledge. But even if our current systems cannot decide issues of this delicacy very accurately, this is no reason to shy away from this part of the task. It is fairly easy to start with a baseline system which believes the reports of major news organizations and government entities, and then to extend that to other cases. For real use of local textual inference technology, source reliability assessment is needed not only for embedded reports as in these examples, but also for evaluating the reliability of the source of the text at the root level.

I do not mean to suggest here that theories of logical semantics have an unsound basis. Both of the example classes presented above can be justified as reasonable cases of particularized conversational implicatures. What they do show is that, if you exclude particularized conversational implicatures from the domain of the textual inference task, what you are left with is likely to be of limited use for practical, real-world applications and no longer corresponds well with human judgements of what texts say. I develop these remarks in the following subsection.

The semantics/pragmatics distinction, or the place of particularized conversational implicature

Following the standard approach within linguistic semantics, ZKC divide inferences into three types, (logical) entailments, conventional implicatures (including presuppositions), and conversational implicatures. The first class captures the literal meaning of the text, while the latter two classes are an attempt to capture how speaker meaning may extend beyond or even just be different from literal meaning. ZKC accept examples where the literal meaning of the text establishes the hypothesis and cases where the hypothesis is a conventional implicature of the text.[9] The more controversial issues occur in the third category of conversational implicatures.

---

[8] A better reason for rejecting it is that the text does not say that the two people are from the same company.

[9] Conventional implicature covers a small class of cases such as deducing from Even Bill stayed out late that most other members of a contextually invoked group are more likely to stay out late than Bill.

The philosopher H.P. Grice distinguished two categories of conversational implicatures: generalized and particularized. Generalized implicatures are common forms of reasoning such as *most X are Y* has an implicature of *not all X are Y*. Particularized implicatures exploit the knowledge of particular situations and utterances. A famous example Grice gives is of a recommendation letter. If the letter says that *Jones has beautiful handwriting and his English is grammatical* then there is an implicature that those are Jones' best qualities, and he is therefore not very creative, smart, or hardworking. Making particularized implicatures requires world knowledge: one has to know a considerable amount about recommendation letters and how they are written for this implicature to arise.

ZKC adopt this distinction and suggest that, with respect to PASCAL RTE, particularized conversational implicatures not be included in "the ways in which an author can be held responsible for her writings on the basis of text internal elements" (p. 34), because too little context is given for them to be reliably calculated. Someone not steeped in the linguistic semantics and pragmatics literature would probably miss the significance of this passage, but the upshot is to argue for excluding from the PASCAL RTE task much of speaker meaning, even though speaker meaning is much closer to the common person's notion of meaning than the literal meaning studied by formal semanticists. This is the wrong direction to head!

ZKC's restriction would again undermine the prospects for having a textual inference task with an interesting interaction between NLP and KR&R. The main subclasses of conventional implicature and generalized conversational implicatures have been codified and could essentially be generated by rule as part of linguistic processing. But there are no algorithms that can calculate the particularized conversational implicatures of a sentence. Rather, it turns on context, world knowledge, and social conventions in varied ways, which require the hearer to deduce the intended meaning of the speaker. Here, there lies an opportunity for KR&R to calculate the likely implicatures in a context. The theory of what conversational implicatures arise is sufficiently fuzzy that many of the inferences that ZKC do not permit could reasonably be viewed as inferences that could be drawn as conversational implicatures. In particular, in the al Qaeda example, if the speaker doubted the veracity of the al Qaeda claim, it would be a violation of Grice's maxim of quantity to not defeat the implicature that the claim is true by adding a phrase such as "but this hasn't been confirmed by local authorities". Without this qualification it is reasonable to accept that the speaker also believes the claim to be true.

The above taxonomy of textual inference types was developed in what Horn (2005) refers to as "the Golden Age of Pure Pragmatics" (roughly, 1965–1985). 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). Horn (2005) essentially argues for the classical account (but even he abandons the traditional account of scalar implicatures, as we will discuss in the next section). Many abandon rather more. Bach (1999) regards the existence of a distinguished class of conventional implicatures as a myth. Others have questioned whether there is a clear distinction between generalized and particularized conversational implicatures. Recanati (2004) rejects the whole idea that a sentence has a literal meaning, emphasizing the essential context dependence of what is said, while still distinguishing this from further things that are implicated.[10] Interestingly, Recanati (2004: 19) proposes regarding what is said as "what a *normal interpreter* would understand as being said, in the context at hand." PASCAL RTE could be viewed as operationalizing such a criterion. Finally, (Szabó 2005) presents contributions from a range of authors, further illustrating the diverse positions currently under discussion. It is not my intention to fully present, let alone to advance, these philosophical discussions, but simply to establish: *(i)* that one should be cautious about basing one's whole research program on a particular viewpoint within

[10] To give just an inkling of the issues, Recanati discusses examples such as a mother, on seeing her son's grazed knee, saying *You're not going to die*. According to the traditional account, the literal meaning of this sentence is that the hearer is immortal, and the much more restricted claim that the son is not going to die *because of his grazed knee* has to be generated as some form of particularized conversational implicature which *overwrites* the literal meaning. Recanati instead argues that what is said is unavoidably context-dependent, and is "You are not going to die from your grazed knee" and this is to be distinguished from implicatures, such as here, perhaps, that the son is being a cry-baby. In cases like this, it is easy to see merit in Recanati's position.

this spectrum, and *(ii)* that adopting the intuitions of a *normal interpreter* as to what is said by a text, as PASCAL RTE does, is quite a defensible position.

THE DATA USED FOR ASSESSING ROBUST TEXTUAL INFERENCE

Is the quality of the Pascal RTE1 data low enough to be problematic?

ZKC emphasize problems with the PASCAL RTE1 data, both errors and things that they think should not be there. An initial issue they note is that "a number of Pascal examples are based on spelling variants or even spelling mistakes … we think they do not belong in a textual inference test bed" (p.34). From the lofty heights of semantic theorizing, such issues may seem distant and mundane. But for those of us actually trying to build systems that do question answering or robust textual inference, we know full well that dealing with issues of matching and normalization is *essential* to getting systems that work at all. *If one wants to build robust working applications or to test operational utility, these are crucial issues to test in an evaluation.* Besides, it is actually easier to normalize *U.S.A.* and *USA* than to deal with most forms of conversational implicature, and one may as well have a system that can handle easy cases of textual inference before trying to work on the hard problems.

What then of the examples that ZKC view as erroneous? We have already discussed above one example (RTE1 ID: 775) which on balance I think is not erroneous. And there are other cases where I would also disagree. ZKC find the following example problematic:

> *Text:* The country's largest private 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 FOLLOWS. (RTE1 ID: 85)

I am rather at a loss to understand why. I think it may be because it requires a certain amount of world knowledge to understand that being kept out of management because they were women is a form of sexual discrimination. But surely precisely what we want to aim towards is building systems that *can understand* this kind of thing.

Like any annotated language resource – but perhaps especially here where it was assembled quickly by graduate students – there are a few items that appear to be just wrong. ZKC discuss:

> *Text:* Green cards are becoming more difficult to obtain.
>
> *Hypothesis:* Green card is now difficult to receive. FOLLOWS. (RTE1 ID: 62)

As they note, an increase in difficulty need not make something absolutely difficult: it is more difficult to get on a plane in the U.S. than 5 years ago, but most people would not judge it as absolutely difficult. However, the number of controversial or wrong examples is relatively few.

As Dagan et al. (2005:5) discuss, at least 3 groups have independently done annotation of portions of the RTE1 data. Groups from the University of Edinburgh, Microsoft, and Mitre report agreement levels with the gold standard data of 95% on all the test set, 96% on one third of the test set, and 91% on one eighth of the development set. Given the differences in the amount of data evaluated and the choice of data set, it seems reasonable to estimate agreement on the RTE1 test set at 95%. While slightly below the agreement level achieved on the carefully defined MUC named entity recognition task (97%) this is well above the < 90% agreement level achieved on various biological named entity recognition tasks (see Dingare et al. 2004 for references). Further, this agreement level gives a Kappa statistic (Carletta 1996) of 0.9. While the Kappa statistic should perhaps be a bit less popular in computational linguistics than it currently is,[11] this is much above the 0.8 level taken to be indicative of "good reliability" in much of the social science literature. This agreement level was in part achieved somewhat artificially by the organizers discarding from their

[11] Inter alia, see *http://ourworld.compuserve.com/homepages/jsuebersax/kappa.htm* and *http://homepages.inf.ed.ac.uk/jeanc/krippendorff.txt*

data set controversial examples, but it is hard to argue that the quality of the data used was sufficiently low as to cause trouble for those attempting to pursue research in robust textual inference.

Is the quality of the PARC (ZKC) KBEval data better?

As a final comparison point for data quality, within the ARDA AQUAINT Knowledge Based Inference pilot (KBEval), groups assembled small selections of development data using a more complex annotation scheme. A group from PARC, including the authors of ZKC, produced a set of 76 examples. Despite the fact that this is an order of magnitude less data, that these items are constructed data, and that the items are much simpler and more parallel than PASCAL RTE data, I agreed with their judgement on only 74 of the 76 examples (97%). This agreement ratio is not statistically significantly higher than that found on the PASCAL RTE1 test set (Fisher's exact test; $p > 0.1$).

For reference, I present the examples on which I disagreed. Firstly:

> *Passage:* Jones arrived in Paris in September last year.
>
> *Question:* Did Jones arrive in Paris in September?
>
> *Answer:* don't know; Polarity: true; Force: strict; Source: linguistic
>
> *Because:* If the questions is asked in September, the answer would have to be no, he arrived last year in September. "in September" send us to the closest one, which would then be interpreted as "this September". (KBEval ID: PARC-12)

Here the annotator appears to have had in mind some very particular context of evaluation. But I would argue that this is not a reasonable default judgement. Consider a scenario like this: "Let's reconstruct the facts. There was an upsurge in terrorist activity last fall. Did Bremer arrive in Baghdad in September?" This question would be satisfied by a document saying "Bremer arrived in Baghdad in September last year." – certainly under the PASCAL assumption (also used in KBEval) that you should assume that identical terms (here September) have identical reference.

The second example is more interesting with respect to categories of inference:

> *Passage:* The man had $20 in his pocket.
>
> *Question:* Did the man have $10 in his pocket?
>
> *Answer:* yes; Polarity: true; Force: strict; Source: linguistic (KBEval ID: PARC-76)

PARC's example is attempting to exploit the classic scalar implicature account of numeric quantities. Under this account, $10 has a meaning (semantics) of "at least $10", and a sense of "exactly $10" arises from a generalized conversational implicature following Grice's Maxim of Quantity. The kind of evidence used in the linguistic literature to support this argument is that if you negate the sentence to "The man didn't have $10 in his pocket", it is the suggested meaning above that appears to be negated giving "he had < $10", rather than negating "exactly $10", which would mean that the negation was true when he had, say, $200 in his pocket. Nevertheless, clearly the person on the street's answer to this question would be: "No, he had $20." The adequacy of the ≥ semantics for numbers suggested by Grice and developed as the theory of scalar implicatures is actually quite controversial in the literature (see, inter alia, Horn (1992), who basically abandons it). Among the problems for this account, one finds in other cases that numbers appear to have an upper bound reading rather than a lower bound reading, for example in the sentence:

> I can fit 3 people in my car.

This shows that the correct interpretation needs to be regarded as a context-dependent particularized semantic implicature, which again indicates the role of common sense knowledge and context in meaning determination. But leaving these issues aside, I believe my answer is better than the proposed answer, because the proposed answer profoundly violates Grice's Maxim of Quantity by giving an answer that is not suitably informative. Hence it is not felicitous, and one should say "no".

Despite the above disagreements, agreement rates should be higher for simple, constructed sentences than for real text. Indeed, recently we have done a reannotation of all of the KBEval data

with two separate annotators, and their agreement rate on the PARC data was noticeably higher than for other data subsets, which used more complex linguistic material drawn from real texts. But the fact that independent annotators achieved similar agreement levels on the PASCAL RTE data shows that the quality of the PASCAL data is good.

Is the informality of the criterion a problem?

Would the task be improved by adopting a more formal definition of entailment rather than relying on the intuitions of an intelligent reader? Interestingly, at Microsoft Research, Dolan et al. (2004, 2005) have developed a corpus of pairs of news sentences judged as paraphrases or non-paraphrases (see also Radev et al. (2003) for a related, but more refined, classification scheme in the context of multidocument summarization). Their goals and standards for annotation are different from PASCAL RTE. They want pairs of sentences that have the same major bits of information, in both directions, but willingly allow minor differences on the details included or whether attribution is present. As they point out, if one wants to have non-trivial instances of two-way equivalence using purely found rather than constructed materials, then it is practically almost a necessity to use a notion of "largely equivalent". But, despite the differences, their experiences with interannotator agreement are very telling and relevant to the PASCAL RTE case as well. Dolan et al. (2005) write:

> "Some specific rating criteria are included in a tagging specification (Section 3), but by and large the degree of mismatch allowed before the pair was judged 'non-equivalent' was left to the discretion of the individual rater: did a particular set of asymmetries alter the meanings of the sentences enough that they couldn't be considered 'the same' in meaning? This task was ill-defined enough that we were surprised at how high interrater agreement was (averaging 83%).
>
> "A series of experiments aimed at making the judging task more concrete resulted in uniformly degraded interrater agreement. Providing a checkbox to allow judges to specify that one sentence entailed another, for instance, left the raters frustrated and had a negative impact on agreement. Similarly, efforts to identify classes of syntactic alternations that would not count against an 'equivalent' judgment resulted, in most cases, in a collapse in interrater agreement. The relatively few situations where we found firm guidelines of this type to be helpful (e.g. in dealing with anaphora) are included in Section 3."

We thus have the apparently paradoxical outcome that an informal task specification leads to quite high interannotator agreement, whereas a formal task specification *lowers* interannotator agreement, often markedly. However, I think there is a good explanation for this. Both the PASCAL RTE task and the Microsoft task are *natural* tasks. They are ones that people engage in and argue over every day ("I had told you that …" [even though I used different words]; "No, you didn't mention the part about …"; …). Judging whether something is a conventional implicature, or should be tagged as a disease in a Named Entity Recognition task is not a natural task, and people tend to perform it much less well, even when presented with a thick rulebook.

The Pascal RTE organizers basically got things right

ZKC suggest as a solution to their issues with the PASCAL RTE1 test set: "Here the test suite is the victim of its self imposed constraints, namely that the relation has to be established between two sentences found in 'real' text. We propose to give up this constraint." (p. 36). Their KBEval pilot data set, discussed above, could perhaps be seen as an instance of the result. ZKC overstate the PASCAL RTE1 organizers' reliance on "real" text. In a good number of instances, I believe that they constructed or adapted a hypothesis so as to test for certain kinds of allowed and false inferences. For instance, in the following example, I highly doubt that they found the hypothesis sentence!

> *Text:* The Osaka World Trade Center is the tallest building in Western Japan.
>
> *Hypothesis:* The Osaka World Trade Center is the tallest building in Japan. FALSE. (RTE1 ID: 2064)

Nevertheless, I believe that the text was always naturally occurring text, and this is to be commended. Not using naturally occurring text would undermine the operational utility of systems that are built for robust textual inference, and would also undermine the scientific goals of the challenge: one wants to empirically examine what types of inferences people make from texts they read and how computers can also come to make them correctly. People have robust and often different intuitions on real data compared to artificial sentences. There is little to be learned about real world, local textual inference from building systems that handle the stylized inference patterns of artificial sentences which are devoid of life.

For the hypothesis, it is highly desirable to derive it from naturally occurring text that is unassociated with the passage text: if someone derives an hypothesis while looking at the passage text, it tends to be much more similar in wording and syntax to the passage text than is typical in real life applications. However, I believe it would be unreasonable to demand that the hypothesis was always unedited naturally occurring text. This would place too high a bar on data production: in many cases it would be impossible to test various issues in system understanding because no appropriate hypothesis text could be found. In particular, it would be impossible to test many *misunderstandings* because there are very limited naturally occurring sources of false text. The hypothesis should be viewed as an experimental probe: in many cases deriving it from naturally occurring data (with or without some editing) will make it a better experimental probe, but in other cases it will be better to construct hypotheses that probe system understanding in particular ways. This does leave to the data set designer a degree of choice as to which things to probe, but the negative effects of this arbitrariness are minimized by grounding the data in real application scenarios and by providing a wide variety of probes. The Microsoft Paraphrase corpus mentioned above shows the limitations of sticking purely to naturally occurring text. Using such a data collection strategy, they note that "insisting on complete sets of bidirectional entailments would have ruled out all but the most trivial sorts of paraphrase relationships, such as sentence pairs differing only [by] a single word or in the presence of titles like 'Mr.' and 'Ms.' ", and so they adopted a rather looser definition of two texts being "more or less semantically equivalent". But to my mind this still allows less interesting and revealing probing of textual inference than is provided by the hypotheses of the PASCAL RTE dataset.

I believe that in all major respects the PASCAL RTE organizers made the right task design choices. The organizers went to considerable effort to source authentic material from representative applications (collecting data from question answering systems, DUC 2004 machine translation data, information extraction tasks, reading comprehension questions, etc.). There have been various suggestions for how to improve on the PASCAL data. But note that none of these suggestions have actually argued that the changes would make the data a better fit to plausible application scenarios. In fact, I believe in most cases that the proposed changes would have the opposite effect. The PASCAL data collection methodology ensured that the text/hypothesis pairs fairly faithfully represent issues that arise in the chosen applications.

A proposed slight change: a text with more context

In the current PASCAL RTE datasets, the text is nearly always one sentence, but it *is* occasionally two sentences.[12] As PASCAL RTE evolves, I would favor having more examples where the text is a short paragraph rather than a single sentence. Firstly, this scenario better matches application scenarios like IR and QA where a retrieval unit of a paragraph is the usual basis for doing textual inference. It also resembles more the Reading Comprehension task with which most people are already familiar. Giving a paragraph of text additionally has the prospect of improving the quality of the data and the confidence people have in it. While doing double reannotation of the KBEval data, I noticed that many of the disagreements in judgement occurred because an annotator attributed some context to the given sentence, but not necessarily the same one as the other annotator. The discussion above of the inference KBEval ID: PARC-12 was a similar case. If more context were provided, I believe that

[12] This latter case has been overlooked by various commentators – but it became very obvious to us when we initially tried to always parse the text as a single sentence!

people's confidence in the answer to an item and interannotator reliability would both increase. That is, there is some justice to people complaining about being quoted out of context. Finally, this change would provide more opportunity to ask questions that involve synthesizing several pieces of information in a paragraph, an important task that is closer to the interests of many KR&R researchers.

REFERENCES

Bach, Kent. 1999. The myth of conventional implicature. *Linguistics and Philosophy* 22:327–366.

Carletta, Jean C. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2), 249–254.

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2005. The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the First PASCAL Challenge Workshop on Recognizing Textual Entailment*. pp. 1–8.

Dingare, Shipra, Malvina Nissim, Jenny Finkel, Christopher Manning, and Claire Grover. 2005. A System For Identifying Named Entities in Biomedical Text: How Results From Two Evaluations Reflect on Both the System and the Evaluations. *Comparative and Functional Genomics* 6: 77–85.

Dolan W. B., C. Quirk, and C. Brockett. 2004. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources. *COLING 2004*. Geneva, Switzerland.

Dolan, Bill, Chris Brockett, and Chris Quirk. 2005. Microsoft Research Paraphrase Corpus. *http://research.microsoft.com/research/nlp/msr_paraphrase.htm*

Dunn, J. Michael and Greg Restall. 2002. Relevance Logic and Entailment. In Dov Gabbay and Franz Guenther (eds), *The Handbook of Philosophical Logic*, 2nd ed. Kluwer. Volume 6, pp. 1–136.

Eco, Umberto. 1984. Dictionary vs. Encyclopedia. In *Semiotics and the Philosophy of Language* (Chapter 2). London: MacMillian.

Haghighi, Aria, Andrew Y. Ng, and Christopher D. Manning. 2005. Robust Textual Inference via Graph Matching. *HLT-EMNLP 2005*, pp. 387-394. *http://nlp.stanford.edu/~manning/papers/rte-emnlp05.pdf*

Harabagiu, S. M.; M. A. Pasca, and S. J. Maiorano. 2000. Experiments with open-domain textual question answering. In *COLING*, 292–298.

Horn, Laurence Robert. 1992. The said and the unsaid. In Chris Barker and David Dowty (eds), *SALT II: Proceedings of the second conference on semantics and linguistic theory*. Columbus: Ohio State University Linguistics Department, pp. 163–192.

Horn, Laurence R. 2005. The Border Wars. In K. Turner and K. von Heusinger (eds.), *Where Semantics Meets Pragmatics*. Elsevier.

Lenat, D. B. and R. V. Guha. 1990. *Building Large Knowledge Based Systems*. Reading, Massachusetts: Addison Wesley.

Moldovan, D. I., C. Clark, S. M. Harabagiu, and S. J. Maiorano. 2003. Cogex: A logic prover for question answering. In *HLT-NAACL*.

Radev, Dragomir, Jahna Otterbacher, and Zhu Zhang. 2003. CSTBank: Cross-document Structure Theory Bank. *http://tangra.si.umich.edu/clair/CSTBank*

Recanati, François. 2004. *Literal Meaning*. Cambridge: Cambridge University Press.

Szabó, Zoltán Gendler. 2005. *Semantics versus Pragmatics*. Clarendon: Oxford University Press.

Wierzbicka, Anna. 1995. Dictionaries vs Encyclopaedias: How to draw the line. In Philip Davis (ed), *Alternative Linguistics: Descriptive and Theoretical Modes*. Amsterdam: John Benjamins. 289-315.

Zaenen, Annie, Lauri Karttunen, and Richard Crouch. 2005. Local Textual Inference: Can it be Defined or Circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pp. 31–36. *http://www.aclweb.org/anthology/W/W05/W05-1206*