Proposed Examples for Pilot Evaluation for Knowledge-Oriented Approaches to Question Answering

PARC Aquaint Team

Palo Alto Research Center
Palo Alto, California 94304, USA

1 Introduction

The examples below are proposed as PARC's contribution to the test suite for the pilot evaluation of Knowledge-Oriented Approaches to Question Answering. The pilot is intended to provide a flexible framework for evaluation and should be useful for knowledge-oriented components as well as for end-to-end systems.

The proposed pilot was inspired by the European PASCAL challenge that introduced textual entailment as a generic evaluation framework for semantic inference in Natural Language Processing. PASCAL defines "textual entailment recognition" as the task of deciding, given two text fragments, whether or not the meaning of one text is entailed by another text given as a premise. For example, the sentence *Rome is located in Lazio province* is marked as a true entailment of the premise *Rome is in Lazio province and Naples is in Campania*. Some AQUAINT projects have participated in the PASCAL challenge, but other researchers have questioned the appropriateness of the PASCAL data and methodology. An underlying objection is that the annotated PASCAL examples lump together in one simple mark-up scheme several different ways in which conclusions might or might not be related to a premise. In some cases there are strict logical/lexical entailments (as in the Rome example), but the relation in other cases is more properly considered as a plausible or likely inference (not a logical entailment) based on unspecified background world knowledge or on general conventions of conversation. Thus, PASCAL's coarse-grained mark-up confuses distinctions that may be important for good question-answering performance. One concern is that the PASCAL challenge may set a low ceiling that will not reward systems that try to make those distinctions. Another is that the PASCAL data may not cover a wide enough range of inference phenomena.

In a preliminary meeting at Stanford and in other discussions at the AQUAINT workshops in Tampa and Palm Springs, there was a general consensus that local textual inference could be the basis for an informative and helpful evaluation of knowledge-oriented approaches to question answering. It was decided that each group involved in this effort would come up with a set of at least 30 sentences for the development suite. What follows is our contribution to the suite.

Our contribution to the test suite concentrates on examples of strict and plausible linguistic (lexical and constructional) inferences. STRICT inferences lead

to conclusions that cannot be cancelled by further assertions in the text, and in that way they differ from PLAUSIBLE inferences. We have limited our contribution to the test suite to inferences that can be drawn purely on the basis of the meanings of words and phrases. But of course the AQUAINT test suite as a whole should also contain examples of inference based on world knowledge that can be assumed to be shared by everybody (e.g. Baghdad is in Iraq).

We adopt a format of source texts followed by possible questions that may or may not be answerable from the local source. We have organized the material in the same way as we did in the white paper proposal on "Local Textual Inference". We changed the format of the inferences to yes-no questions because that is more appropriate for AQUAINT. With each item we give first the source, then a question, an appropriate answer, and a classification along the lines of the scheme we proposed in the white paper. We indicate whether the inference is strict or plausible, and whether it depends purely on linguistic knowledge (as all of ours do) or also involves some degree of background world knowledge.

2 Evaluation Examples

(1) Some students came to school by car.
Did any students come to school?
yes
STRICT, LINGUISTIC

(2) No students came to school by car.
Did any students come to school?
unknown
STRICT, LINGUISTIC

(3) John drove legally.
Did John drive?
yes
STRICT, LINGUISTIC

(4) John drove predictably.
Did John drive?
yes
STRICT, LINGUISTIC

(5) Legally, John could drive.
Did John drive?
unknown
STRICT, LINGUISTIC

(6) Predictably, John drove.
Did John drive?
yes
STRICT, LINGUISTIC

(7) The technician cooled the room.
Did the technician lower the temperature of the room?
yes
STRICT, LINGUISTIC

(8) The technician raised the temperature of the room.
Did the technician cool the room?
no
STRICT, LINGUISTIC

(9) The president visited Iraq in September.
Has the president gone to Iraq?
yes
STRICT, LINGUISTIC

(10) Jones has visited Iraq.
Did Jones visit Iraq in September?
unknown
STRICT, LINGUISTIC

(11) Jones arrived in Paris in September last year.
Did Jones arrive in Paris last year?
yes
STRICT, LINGUISTIC

(12) Jones arrived in Paris in September last year.
Did Jones arrive in Paris in September?
unknown
STRICT, LINGUISTIC

(13) Jones arrived on a Sunday in September.
Did Jones arrive on a Sunday?
yes
STRICT, LINGUISTIC

(14) Jones arrived on a Sunday in September.
Did Jones arrive in September?

yes
STRICT, LINGUISTIC

(15) The president left after the diplomat arrived.
Did the diplomat arrive before the president leave?
yes
STRICT, LINGUISTIC

(16) No US congressman has visited Iraq since the war ended.
Has Jones, a US Congressman, visited Iraq after the war ended?
no
STRICT, LINGUISTIC

(17) No US congressman has visited Iraq since the war.
Did Jones, a US Congressman, visit Iraq before the war?
unknown
STRICT, LINGUISTIC

(18) No US congressman visited Iraq until the war.
Did any US congressman visit Iraq before the war?
no
STRICT, LINGUISTIC

(19) Some students arrived at the school on Sunday.
Were there any students at the school on Sunday?
yes
STRICT, LINGUISTIC

(20) No students arrived at the school on Sunday.
Were there any students at the school on Sunday?
unknown
STRICT, LINGUISTIC

(21) There were no students at the school on Sunday.
Did any students arrive at the school on Sunday?
no
STRICT, LINGUISTIC

(22) The diplomat left Baghdad last week.
Has the diplomat been to Baghdad?
yes
STRICT, LINGUISTIC

(23) The diplomat will arrive in Baghdad next week.
Has the diplomat been to Baghdad?
unknown
STRICT, LINGUISTIC

(24) The president knows that the diplomat left Baghdad.
Has the diplomat been to Baghdad?
yes
STRICT, LINGUISTIC

(25) The president hasn't gone to Iraq since the diplomat left Baghdad.
Has the diplomat been to Baghdad?
yes
STRICT, LINGUISTIC

(26) The president hasn't gone to Iraq since the diplomat left Baghdad.
Has the president been to Iraq?
unknown
STRICT, LINGUISTIC

(27) The diplomat didn't manage to leave Baghdad.
Has the diplomat been to Baghdad?
yes
STRICT, LINGUISTIC

(28) The diplomat hasn't managed to leave Baghdad.
Is the diplomat in Baghdad now?
yes
STRICT, LINGUISTIC

(29) The room was full of intelligent women.
Was the room full of women?
yes
STRICT, LINGUISTIC
(See appendix: Simple affirmative sentences are generally UPWARD MONOTONIC.[1])

(30) The room was full of women.
Was the room full of intelligent women?
unknown

[1] A sentence is upward monotonic iff it remains true when it is broadened, e.g. by replacing *intelligent women* by the more general term *women*.

STRICT, LINGUISTIC

(31) Children are not admitted to the theatre.
Are small children admitted to the theatre?
no
STRICT, LINGUISTIC
(See appendix: Simple negative sentences are generally DOWNWARD MONOTONIC.[2])

(32) Small children are not admitted to the theatre.
Are children admitted to the theatre?
unknown
STRICT, LINGUISTIC

(33) All companies have to file annual reports.
Do all Fortune 500 companies have to file annual reports?
yes
STRICT, LINGUISTIC
(See appendix: *All* is downward monotonic with respect to its RESTRICTOR.[3])

(34) All Fortune 500 companies have to file annual reports.
Do all companies have to file annual reports?
unknown
STRICT, LINGUISTIC
(See appendix: *All* is not upward monotonic with respect to its restrictor.)

(35) All companies have to file annual reports to the SEC.
Do all companies have to file annual reports?
yes
STRICT, LINGUISTIC
(See appendix: *All* is upward monotonic with respect to its SCOPE.[4])

(36) All companies have to file annual reports.
Do all companies have to file annual reports to the SEC.
unknown

[2] A sentence is downward monotonic iff it remains true when it is narrowed, e.g. by replacing *children* by the more specific term *small children*.
[3] A quantifier Q is downward monotonic with respect to its restrictor ϕ iff $((Q\ \phi)\ \psi)$ remains true when ϕ is narrowed, e.g. from *companies* to *Fortune 500 companies*.
[4] A quantifier Q is upward monotonic with respect to its scope ψ iff $((Q\ \phi)\ \psi)$ remains true when ψ is broadened, e.g. from *have to file reports to the* SCE to just *have to file reports*.

STRICT, LINGUISTIC
(See appendix: *All* is not downward monotonic with respect to its scope.)

(37) No delegates finished the report.
Did any delegate finish the report on time?
no
STRICT, LINGUISTIC
(*No* is downward monotonic with respect to its scope.)

(38) The US troops stayed in Iraq although the war was over.
Was the war over?
yes
STRICT, LINGUISTIC

(39) Since it was cold, he closed the window.
Was it cold?
yes
STRICT, LINGUISTIC

(40) John didn't visit us after he returned from Spain.
Did John return from Spain?
yes
STRICT LINGUISTIC

(41) Hanssen, who sold FBI secrets to the Russians, could face the death penalty.
Did Hanssen sell FBI *secrets to the Russians?*
yes
STRICT, LINGUISTIC

(42) The New York Times reported that Hanssen, who sold FBI secrets to the Russians, could face the death penalty.
Did Hanssen sell FBI *secrets to the Russians?*
yes
STRICT, LINGUISTIC

(43) The New York Times reported that Hanssen sold FBI secrets to the Russians and could face the death penalty.
Did Hanssen sell FBI *secrets to the Russians?*
unknown
STRICT, LINGUISTIC

(44) Bush said that it was Khan who sold centrifuges to North Korea.
Were centrifuges sold to North Korea?
yes
STRICT, LINGUISTIC

(45) Bush said that Khan sold centrifuges to North Korea.
Were centrifuges sold to North Korea?
unknown
STRICT, LINGUISTIC

(46) What we found in Iraq was rusted shrapnel.
Did we find anything in Iraq?
yes
STRICT, LINGUISTIC

(47) The fact that Bin Laden was in Tora Bora lead to the suspicion that the Afghan campaign was mismanaged.
Was Bin Laden in Tora Bora?
yes
STRICT, LINGUISTIC

(48) The fact that Bin Laden was in Tora Bora lead to the suspicion that the Afghan campaign was mismanaged.
Was the Afghan campaign mismanaged?
unknown
STRICT, LINGUISTIC

(49) The paper concluded that the election had been rigged.
Was the election rigged?
unknown
STRICT, LINGUISTIC

(50) Ames was, as the press reported, a successful spy.
Was Ames a successful spy?
yes
STRICT, LINGUISTIC

(51) The press reported that Ames was a successful spy.
Was Ames a successful spy?
unknown
STRICT, LINGUISTIC

(52) The US forgot that the Afghans speak several different languages.
Do the Afghans speak several different languages?
yes
STRICT, LINGUISTIC

(53) Bush realized that the US Army had to be transformed to meet new threats.
Did the US Army have to be transformed to meet new threats?
yes
STRICT, LINGUISTIC

(54) Bush didn't realize that Afghanistan is land-locked.
Is Afghanistan land-locked?
yes
STRICT, LINGUISTIC

(55) There is a belief that the US will invade Syria.
Will the US invade Syria?
unknown
STRICT, LINGUISTIC

(56) It is not surprising that Bush has the lead in Ohio.
Does Bush have the lead in Ohio?
yes
STRICT, LINGUISTIC

(57) It is not likely that Bush has the lead in Ohio.
Does Bush have the lead in Ohio?
unknown
STRICT, LINGUISTIC

(58) Kerry knew that Edwards would accept the nomination.
Did Kerry know whether Edwards would accept the nomination?
yes
STRICT, LINGUISTIC

(59) Tom knows that Naples is in Campania.
Does Tom know where Naples is?
yes
STRICT, LINGUISTIC

(60) We met in September during the feast.

Did the feast take place in September?
yes
STRICT, LINGUISTIC

(61) It is false that Bin Laden was seen in Tora Bora.
Was Bin Laden seen in Tora Bora?
no
STRICT, LINGUISTIC

(62) It follows that Bin Laden was in Tora Bora.
Was Bin Laden in Tora Bora?
yes
STRICT, LINGUISTIC

(63) It is likely that Bin Laden was in Tora Bora.
Was Bin Laden in Tora Bora?
unknown
STRICT, LINGUISTIC

(64) Tony Hall left Amman on Sunday.
Was Tony Hall in Amman on Sunday?
yes
STRICT, LINGUISTIC

(65) Tony Hall left Amman on Sunday.
Was Tony Hall in Amman on Saturday?
unknown
STRICT, LINGUISTIC

(66) Khan sold 10 centrifuges to North Korea.
Did North Korea buy 10 centrifuges?
yes
STRICT, LINGUISTIC

(67) The US invasion of Afghanistan prevented Al-Qaida from attacking Ryad in 2002.
Did Al-Qaida attack Ryad in 2002?
no
STRICT, LINGUISTIC

(68) The administration managed to track down the perpetrators.
Did the administration track down the perpetrators?

yes
STRICT, LINGUISTIC

(69) The administration didn't manage to track down the perpetrators.
Did the administration track down the perpetrators?
no
STRICT, LINGUISTIC
(See appendix: *Manage* is an IMPLICATIVE verb.[5])

(70) Bush didn't have the time to read the report.
Did Bush read the report?
no
STRICT, LINGUISTIC

(71) Bush had the time to read the report.
Did Bush read the report?
yes
PLAUSIBLE, LINGUISTIC
(See appendix: *have the time* is a SEMI-IMPLICATIVE construction.[6])

(72) The president wasn't able to attend the meeting.
Did the president attend the meeting?
no
STRICT, LINGUISTIC

(73) The president was able to attend to meeting.
Did the president attend the meeting?
yes
PLAUSIBLE, LINGUISTIC
(See appendix: The implicature can be strengthened by adverbs such as *luckily*.)

(74) Many soldiers were killed in the ambush.
Were all soldiers killed in the ambush?
no
PLAUSIBLE, LINGUISTIC

[5] Implicative constructions such as *manage/fail/happen to ...*, etc. yield an entailment both in affirmative and negative sentences.

[6] Semi-implicative constructions such as *have the time/be able to ...* yield an entailment in negative sentences and a conversational implicature in affirmative sentences. Conversational implicatures are CANCELLABLE. It is not a contradiction to say *Bush had the time to read the report but he did not bother to read it.*

(75) The man had $20 in his pocket.
Did the man have $40 in his pocket?
no
PLAUSIBLE, LINGUISTIC

(76) The man had $20 in his pocket.
Did the man have $10 in his pocket?
yes
STRICT, LINGUISTIC

3 Appendix: Selection of Phenomena

As stated at the beginning, a concern with PASCAL was the limited range of examples covered. Our contribution to the test suite deliberately aims to broaden the range of linguistically significant phenomena dealt with. In this appendix, we briefly set out some of the considerations that went into the selection of examples, and provide some explanation for the terminology used in comments on some examples. Our test suite concentrates on examples of logical and linguistic inferences of three types: entailments, conventional implicatures and conversational implicatures.

Textual inferences mainly arise from the lexical items that are used in the assertions made in the text. Several types of lexical items, such as quantifiers and negation, play a privileged role in entailments as can be seen from the examples above, but all lexical classes lead to entailments. Of special interest are monotone upward or downward entailments. The fact that both kinds exist means that simple matching on an inclusion of relevant material cannot not work as a technique to detect entailments. Upward monotone expressions preserve truth by leaving out material whereas downward monotone expressions don't; instead, adding material to them can be truth preserving.[7]

Conventional implicatures in general encode material that the author wants to present as not at issue, as not controversial. This material is presented as a starting point for the assertions that the author wants to make or because she wants to remind the reader of something he might not have in mind or tell him something he should know but might not. This not-at-issue material in text such as newspapers is especially interesting for problems like question answering. It tends to be less controversial than the foregrounded material: it expresses what the author is committed to or assumes to be commonly agreed-upon knowledge, whereas the new material is often attributed to other sources without the author taking responsibility for it or presenting it as true ([2] [3]). PASCAL implicitly recognizes the importance of this type of material by the number of appositives, one kind of conventional implicature, that are part of its test suite.

[7] Dagan and Glickman [1] explore inferencing directly over lexical-syntactic representation but they only consider upward monotonic expressions.

Conversational implicatures are legitimate inferences that are plausible rather than strict, and can be cancelled by further assertions in the text. Conversational implicatures rest on the assumption that the author is trying to impart information and not trying to mislead: he will give all the information he deems necessary and not more([4] [5]). A report saying that *Many soldiers were killed in the ambush* implies that some soldiers survived but does not logically exclude the possibility that they all died. This implication can be cancelled, for example, by a subsequent assertion in the report indicating that in fact there were no survivors.

The test suite annotation does not directly distinguish between simple entailments, conventional implicatures, or conversational implicatures. Entailments and conventional implicatures are both catalogued as STRICT, LINGUISTIC. Conversational implicatures are catalogued as PLAUSIBLE, LINGUISTIC.

References

1. Dagan, I., Glickman, O.: Probabilistic textual entailment: Generic applied modeling of language variablity. In: Learning Methods for Text Understanding and Mining, Grenoble (2004)
2. 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
3. Potts, C.: The Logic of Conventional Implicatures. University of California, Santa Cruz, CA (2003) Ph.D dissertation, Department of Linguistics.
4. Grice, P.H.: Studies in the Way of Words. Harvard University, Cambridge, MA (1989)
5. Horn, L.: Implicature. In Horn, Ward, eds.: Handbook of Pragmatics. Blackwell, Oxford (2003)