Local Textual Inference: can it be defined or circumscribed?

Annie Zaenen
Palo Alto Research Center
3333, Coyote Hill Road
Palo Alto, CA 94304
zaenen@parc.com

Lauri Karttunen
Palo Alto Research Center
3333, Coyote Hill Road
Palo Alto, CA 94304
karttunen@parc.com

Richard Crouch
Palo Alto Research Center
3333, Coyote Hill Road
Palo Alto, CA 94304
crouch@parc.com

Abstract

This paper argues that local textual inferences come in three well-defined varieties (entailments, conventional implicatures/presuppositions, and conversational implicatures) and one less clearly defined one, generally available world knowledge. Based on this taxonomy, it discusses some of the examples in the PASCAL text suite and shows that these examples do not fall into any of them. It proposes to enlarge the test suite with examples that are more directly related to the inference patterns discussed.

1 Introduction

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. This 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. The task also shies away from creative metaphorical or metonymic use of language and makes the assumption that referential assignments remain constant for entities that are described in the same way. These all seem good features of the proposal as it stands.

Looking at the challenge as it was put before the community, however, we feel that it might be useful to try to circumscribe more precisely what exactly should count as linguistic knowledge. In this paper we make a stab at this in the hope of getting a discussion going. For reasons that will become clear, we prefer to talk about TEXTUAL INFERENCES rather than about textual entailments when referring to the general enterprise. We first explicitate what we think should be covered by the term textual inferences, we then look at the PASCAL development suite in the light of our discussion and we conclude with a short proposal for extensions to the test suite.

Before even starting at this, a point of clarification needs to be made: the correspondence of a linguistic object to an object in the real world goes beyond what can be learned from the text itself. When somebody says or writes *The earth is flat* or *The king of France is bald* because (s)he is a liar or ill-informed, nothing in these linguistic expressions in themselves alerts us to the fact that they do not correspond to situations in the real world (we leave texts in which the author signals consciously or unconsiously that he is lying or fibbing out of consideration here.) What the text does is give us information about the stance its author takes vis-à-vis the events or states described.

It is thus useful to distinguish between two ingredients that go into determining the truth value of an utterance, one is the trustworthiness of the utterer and the other is the stance of the utterer vis-à-vis the truth of the content. The latter we will call the veridicity of the content. When we talk about textual inferences we are only interested in veridicity not in the truth which lies beyond what can be inferred from texts. Or, maybe more realistically, we assume a trustworthy author so that veridical statements are also true.

2 Varieties of local textual inferences

Under this assumption of trustworthiness, semantics and pragmatics as practiced by philosophers and linguists can give us some insights that are of practical relevance. Work done in the last century has led researchers to distinguish between entailments, conventional implicatures and conversational implicatures. We describe these three classes of inferences and illustrate why the distinctions are important for NLP.

2.1 Entailments

The most uncontroversional textual inferences are those that can be made on the basis of what is asserted in a text. If the author makes the statement that *Tony Hall arrived in Baghdad on Sunday night*, then we can conclude that *Tony Hall was in Baghdad on Sunday night* (keeping referring expressions constant, as proposed in the PASCAL task). The second sentence is true when the first is true (assuming we are talking about the same Tony Hall, the same Baghdad and the same Sunday) just by virtue of what the words mean.

In simple examples such as that in (1)

(1) Bill murdered John.
Bill killed John.

one can go to a resource such as WordNet, look up *murder*, discover that it means *kill* with some further conditions. "Ontologies" or thesauruses typically order terms in a hierarchy that encodes a relation from less specific at the top of the hierarchy to more specific at the bottom. In simple clauses the replacement of a more specific term with a less specific one, ensures an upward monotonic relation between these sentences. As is well known this relation is inversed when the sentences are negated.[1]

(2) Bill didn't murder John.
does not entail *Bill didn't kill John.*

but

(3) Bill didn't kill John.
does entail *Bill didn't murder John.*

Monotonicity relations also hold when adjectival modification is introduced as in (4)

(4) Ames was a clever spy.
entails *Ames was a spy.*

Again negation reverses the entailment:

(5) Ames wasn't a spy.
entails *Ames wasn't a clever spy.*

Quantifiers, easily among the most intensively studied lexical items, also exhibit upward or downward monotonicity.[2] To give just one example:

(6) All companies have to file annual reports.
entails *All Fortune 500 companies have to file annual reports.*

but

(7) All companies have to file annual reports.
does not entail *All companies have to file annual reports to the* SEC.

The fact that there are both upwards monotonic and downwards monotonic expressions means that simple matching on an inclusion of relevant material cannot work as a technique to detect entailments. Upward monotone expressions preserve truth by leaving out material whereas downward monotone expressions don't: adding material to them can be truth preserving.[3]

Apart from a more specific/less specific relation, lexical items can establish a part-subpart relation between the events they describe. If we followed the first sentence in (1) by

(8) John died.

we would still have a lexical inference. In this case one in which the event described in the second sentence is a subpart of the event described in the first.

The investigation of entailments leads one to distinguish several types of lexical items that have predictable effects on meaning that can be exploited to discover sentences that are inferentially related (by real entailments in this case). Other examples are scope bearing elements (an aspect of meaning that often leads to ambiguities which are not always easily perceived) and perception reports.

[1] A sentence is downward monotonic iff it remains true when it is narrowed. A sentence is upward monotonic when it remains true when it is broadened.

[2] A quantifier *Q* is downward monotonic with respect to its restrictor ϕ iff *((Q ϕ) ψ)* remains true when the ϕ is narrowed, e.g. from *companies* to *Fortune 500 companies*. A quantifier *Q* is upward monotonic with respect to its scope ψ iff *((Q ϕ) ψ)* remains true when ψ is broadened, e.g. from *have to file reports to the* SCE to just *have to file reports*.

[3] Dagan and Glickman (2004) explore inferencing by syntactic pattern matching techniques but consider only upward monotonic expressions. Their proposal ensures loss of recall on downward monotonic expressions.

Two types of relations deserve special mention here because they are pervasive and they are at the borderline between linguistic and world knowledge: temporal relations and spatial relations. Whether knowing that Tuesday follows Monday or that there are leap years and non-leap years 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. The consequences in a Eucledian space of the place and movement of objects are similar. There is a rich set of entailment relations that builds on these temporal and spatial notions.

2.2 Conventional Implicatures[4]

Apart from making assertions, however, an author will often "conventionally implicate" certain things. We use here the term conventional implicature for what has been called by that name or labeled as (semantic) presupposition. Some of us have argued elsewhere there is no need for a distinction between these two notions (Karttunen and Peters, 1979) and that presupposition is a less felicitous term because it tends to be confused with "old information".

Traditionally these implications are not considered to be part of what makes the sentence true, but the author is COMMITTED to them and we consider them part of what textual inferences should be based on. We take this position because we think it is reasonable, for IE tasks, to assume that material that is conventionally implicated can be used in the same way as assertions, for instance, to provide answers to questions. When somebody says *Bill acknowledges that the earth is round*, we know something about the author's as well as Bill's beliefs in the matter, namely that the author is committed to the belief that the earth is round.

If all conventionally implied material were also discourse old information, this might not matter very much as the same information would be available elsewhere in the text, but often conventionally implied material is new information that is presented as not being under discussion. Conventional implicatures are a rich source of information for IE tasks because the material presented in them is supposed to be non-controversial. In newspapers and other information sources they are a favorite way to distinguish background knowledge, that the reader might have or not, without confusing it with what is newsworthy in the report at hand. A very common example of this, exploited in the PASCAL test suite, is the use of appositives. illustrated in the following example:

(9) The New York Times reported that Hanssen, who sold FBI secrets to the Russians, could face the death penalty.
Did Hanssen sell FBI reports to the Russians?
YES

From the perspective of IE tasks, the way conventional implicatures behave under negation is one reason to pay close attention to them. The following examples illustrate this:

(10) Kerry realized that Bush was right.
Bush was right.

(11) Kerry didn't realize that Bush was right.
Bush was right.

Other types of embedded clauses that are conventionally implicated are temporal adverbials (except those introduced by *before* or *until*. Other types of material that can introduce a conventional implicature are adverbial expressions such as *evidently* and simple adverbs such as *again* or *still*.

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.

2.3 Conversational Implicatures[5]

Authors can be held responsible for more than just assertions and conventional implicatures. Conversational implicatures are another type of author commitment. A conversational implicature rests on the assumption that, in absence of evidence to the contrary, a collaborative author will say as much as she

[4]For more on conventional implicatures, see e.g. Karttunen and Peters (1979) and Potts (2005)

[5]For more on conversational implicatures, see e.g. Grice (1989) and Horn (2003)

knows. So if Sue says that she has four children, we tend to conclude that she has no more than four. This type of implicature can be destroyed without any contradiction arising: *He not only ate some of the cake, he ate all of it.* Within the context of a textual inference task such as that defined in the PASCAL initiative, it is clear that inferences based on conversational implicatures might be wrong: PASCAL doesn't give the context. In a more developed type of inference task, a distinction should be made between this type of inference and the ones we discussed earlier, but when inferencing is reduced to one sentence it seems more reasonable to take generalized conversational implicatures into account as bona fide cases of inferences (except of course if they are cancelled in the sentence itself, as in the example above).

(14) I had the time to read your paper.

conversationally implies that I read your paper. But it could be followed by *but I decided to go play tennis instead.*

(15) Some soldiers were killed.

conversationally implies *Not all soldiers were killed.* But it could be cancelled by *In fact we fear that all of them are dead.*

(16) He certainly has three children.

conversationally implies *He doesn't have more than three children* but it could be followed by *In fact he has five, three daughters and two sons.*

Apart from the general conversational implicatures, implicatures can also arise by virtue of something being said or not said in a particular context. If in a letter of recommendation, one praises the candidate's handwriting without saying anything about his intellectual abilities, this allows the reader to draw some conclusions. We assume here that this type of inference is not part of the PASCAL task, as too little context is given for it to be reliably calculated.

One might agree with the analysis of various sources of author commitment given above but be of the opinion that it doesn't matter because, given enough data, it will come out in the statistical wash. We doubt, however, that this will happen any time soon without some help: the semantic distinctions are rather subtle and knowing about them will help develop adequate features for statistical training.

It might also be thought that the generalizations that we need here can be reduced to syntactic distinctions. We don't have the space to show in great detail that this is not the case but some reflection on and experimentation with the examples given throughout this paper will convince the reader that this is not the cases. For instance, if one replaces the adjective *clever* with the equally good adjective *alleged* in (4) above, the entailment relation between the sentences doesn't hold anymore. Substituting *show* for *realize* in (11) has the same effect.

2.4 Some world knowledge?

In our mind this exhausts the ways in which an author can be held responsible for her writings on the basis of text internal elements. Textual inferences are based on textual material that is either an entailment of what is explicitly asserted, or material that conventionally or conversationally implied by the author. These inferences can be made solely on the basis of the way the meaning of the words and construction she uses are related to other words and constructions in the language. 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: the societies we live in use a common view of time to describe events and rely on the assumptions of Euclidean geometry, leading to shared calendars and measurement systems. It would be impossible to separate these from linguistic knowledge. 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.

3 The PASCAL development suite.

We now discuss some of the PASCAL development set examples in the light of the discussion above and explain why we think some of them do not belong in a textual inference task. First a number of PASCAL examples are based on spelling variants or even spelling mistakes. While it is clear that coping with this type of situation is important for NLP applications we think they do not belong in a textual inference test bed. We first discuss a couple of examples

that we think should not have been in the test suite and then some that do not confirm to our view on inferencing but which might belong in a textual inference test suite.

3.1 Errors?

A problem arises with an example like the following:

(17) A farmer who was in contact with cows suffering from BSE – the so-called mad cow disease – has died from what is regarded as the human form of the disease.
Bovine spongiform encephalopathy is another name for the "mad cow disease".
TRUE

If one googles BSE, one finds that it is an abbreviation that can stand for many things, including the Bombay, Bulgarian, Baku or Bahrain Stock Exchange, Breast Self-Examination, and Brain Surface Extractor. To select the right alternative, one needs the knowledge that "bovine spongiform encephalopathy" is a name of a disease and the other competing BSE expansions are not.

The authors of the PASCAL test suite don't seem to allow for as much world knowledge when they mark the following relation as FALSE.

(18) "I just hope I don't become so blissful I become boring" – Nirvana leader Kurt Cobain said, giving meaning to his "Teen Spirit" coda, a denial.
"Smells Like Teen Spirit" is a song by Nirvana.
FALSE

Apparently, it is NOT OK to know that the Nirvana song "Smells like Teen Spirit" is often referred to as "Teen Spirit". But why should we then know that bovine spongiform encephalopathy is a disease?

The test suite also contains examples that can only be classified as plain errors. A couple of examples are the following:

(19) Green cards are becoming more difficult to obtain.
Green card is now difficult to receive.
TRUE

Something that is becoming more difficult can still be easy, if it starts out that way.

(20) Hippos do come into conflict with people quite often.
Hippopotamus attacks human.
TRUE

For somebody who knows a lot about hippos it might be reasonable to assume that a conflict is necessarily an attack but in general there is no inference: *conflict* is the less general term and *attack* the more specific one.

(21) A statement said to be from al Qaida claimed the terror group had killed one American and kidnapped another in Riyadh.
A U.S. citizen working in Riyadh has been kidnapped.
TRUE

This seems betray a rather implausible belief in the claims of al Qaida and while we are assuming that the author of the text is trustworthy, this assumption does not extend to the sources he invokes. In this case especially, the use of *claim* can be construed as indication the doubt of the author about the veracity of what the source says.

(22) Wal-Mart is being sued by a number of its female employees who claim they were kept out of jobs in management because they were women.
Wal-Mart is sued for sexual discrimination.
TRUE

A minute of reflection will make clear that here the relation between the two sentences involves quite a bit of specialized legal knowledge and goes beyond textual inferencing. How is *sexual discrimination* different from *sexual harassment*?

(23) South Korean's deputy foreign minister says his country won't change its plan to send 3000 soldiers to Iraq.
South Korea continues to send troops.
TRUE

We assume that in context the second sentence means that South Korea continues to plan to send troops but normally *continue* does not mean *continue to plan* and the first sentence certainly doesn't imply that South Korea has already sent troops. Here the way the test suite has been put together leads to odd results. A headline is paired up with a full sentence. Headlines are not meant to be understood completely out of context and it would be prudent to use them sparingly in inference tasks of the sort proposed here. We discuss other consequences of the way the test suite was constructed in the next subsection with examples that to our mind need some kind of accommodation.

3.2 Not a textual inference as such but . . .

There are a couple of examples such as the following in the test suite:

(24) The White House failed to act on the domestic threat from al Qaida prior to September 11, 2001.
White House ignored the threat of attack.
TRUE

Here there is no entailment either way and surely *fail to act* is not a synonym of *ignore*. The examples are due to the way the PASCAL test suite was put together. It was evidently at least in part developed by finding snippets of text that refer to the same event in different news sources; this is a fertile method for finding inferences but it will lead to the inclusion of some material that mixes factual description and various APPRECIATIONS of the described facts. For instance in (24) above, two different authors described what the White house did, putting a different spin on it. While the fact described in both cases was the same, the appreciations that the two renderings give, while both negative, are not equivalent. But although there is no legitimate inference for the sentences as a whole, they both entail that the White House did not act. 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.

Another maybe simpler illustration of the same problem is (25):

(25) The report catalogues 10 missed opportunities.
The report lists 10 missed opportunities.

Although *catalogue* and *list* do not have the same meaning, they may in some cases be used interchangeably because, again, there is a common entailment:

(26) According to the report, there were 10 missed opportunities.

One can conceive of a thesaurus where *catalogue* and *list* would have a low level common hypernym (in WordNet they don't) or a statistically inferred word class that would make the common entailment explicit, but that relation should not be confused with an inference between the two sentences in (25).

4 A proposal for some refinements

As the discussion above has shown, the way the test suite was put together leads sometimes to the inclusion of material that should not be there given the definition of the task. Most of the data that form the basis of PASCAL are extracted from different newspaper articles about the same event, often from the same newswire. This means that the information packaging is very similar, reducing the constructional and lexical range that can be used to express a same idea. This situation will not pertain in the more general setting of question answering and many types of paraphrases or inferences that would be useful for question answering in general will not be found or will be very rare in PASCAL-like suites.

We would propose to augment the types of pairs that one can get through the PASCAL extraction techniques with some that take the type of relations that we have discussed explicitly into account. It can be objected that this introduces a new level of artificiality by allowing made-up sentences but the separation of world knowledge from linguistic knowledge is in any case artificial. But it is necessary because we will not be able to solve the inferencing problem without slicing the task into manageable pieces.

Acknowledgments

This article was supported in part by the Advanced Research and Development Agency (ARDA) within the program for Advanced Question Answering for Intelligence (AQUAINT). Thanks to all the members of PARC's AQUAINT team.

References

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variablity. In *Learning Methods for Text Understanding and Mining*, Grenoble, January.

Paul H. Grice. 1989. *Studies in the Way of Words*. Harvard University, Cambridge, MA.

Larry Horn. 2003. Implicature. In Horn and Ward, editors, *Handbook of Pragmatics*. Blackwell, Oxford.

Lauri Karttunen and Stanley Peters. 1979. Conventional implicature. In Choon-Kyu Oh and David A. Dinneen, editors, *Syntax and Semantics, Volume 11: Presupposition*, pages 1–56. Academic Press, New York.

Christopher Potts. 2005. *The Logic of Conventional Implicatures*. Oxford Studies in Theoretical Linguistics. Oxford University Press, Oxford.