Defining Textual Entailment

Daniel Z. Korman
Department of Philosophy, University of California, Santa Barbara, Santa Barbara, California.
E-mail: dkorman@ucsb.edu

Eric Mack
Independent Scholar, 1177 Avenue of the Americas, Floor 31, New York, NY 10036.
E-mail: Ericalan11@gmail.com

Jacob Jett
Center for Informatics Research in Science and Scholarship, School of Information Sciences, University of Illinois at Urbana-Champaign, 501 East Daniel Street, MC-493, Champaign, IL 61820-6211.
E-mail: jjett2@illinois.edu

Allen H. Renear
School of Information Sciences, University of Illinois at Urbana-Champaign, 501 East Daniel St., MC-493, Champaign, Illinois. E-mail: renear@illinois.edu

Textual entailment is a relationship that obtains between fragments of text when one fragment in some sense implies the other fragment. The automation of textual entailment recognition supports a wide variety of text-based tasks, including information retrieval, information extraction, question answering, text summarization, and machine translation. Much ingenuity has been devoted to developing algorithms for identifying textual entailments, but relatively little to saying what textual entailment actually *is*. This article is a review of the logical and philosophical issues involved in providing an adequate definition of textual entailment. We show that many natural definitions of textual entailment are refuted by counterexamples, including the most widely cited definition of Dagan et al. We then articulate and defend the following revised definition: T textually entails $H =_{df}$ typically, a human reading T would be justified in inferring the proposition expressed by H from the proposition expressed by T. We also show that textual entailment is context-sensitive, nontransitive, and nonmonotonic.

Introduction

Textual entailment is a relationship that obtains between fragments of text when one fragment in some sense implies the other. Recognizing such connections is an important and routine part of linguistic communication, whether in common conversation or scientific literature. As a consequence, the automation of textual entailment recognition can support a wide variety of text-based tasks, including information retrieval, information extraction, question answering, text summarization, and machine translation (Bos & Markert, 2005; Harabagiu & Hickl, 2007; Padó et al., 2009; Blake, 2011). More generally, almost all tools and applications in use today to navigate and exploit online textual material stand to benefit from the automated identification of textual entailments, which can help end users navigate the deluge of information from online communities, open access article databases, and other such sources (Renear & Palmer, 2009). For instance, a question answering application would benefit from recognizing that the text "John bought a novel yesterday" textually entails "John bought a book," enabling it to identify the former as a suitable response to the query "Did John buy a book?"[1]

Textual entailment research largely consists of developing and exploring approaches to algorithmic identification of entailments. Considerable ingenuity has gone into developing such algorithms, particularly under the auspices of the PASCAL project's Recognizing Textual Entailment (RTE) challenges. Building upon the RTE challenges' successes, numerous researchers are now in the midst of testing

Received January 30, 2017; revised September 30, 2017; accepted December 17, 2017

 DOI: 10.1002/asi.24007

[1]Cf. Dagan and Glickman (2004).

systems that leverage textual entailments: pioneering new kinds of information retrieval systems (Udayakumar et al., 2014), developing new means of analyzing text (Magnini et al., 2014; Kolterman et al., 2015) and interpreting metaphors (Mohler et al., 2013), and validating results from question answering workflows (Gómez-Adorno et al., 2013).[2]

By contrast, relatively little effort has gone into determining exactly what textual entailment, the object of study, actually *is*. Since such definitions are the cornerstone of the instructions that annotators use when making gold standard data sets—which in turn are used to develop, train, and test algorithms for recognizing textual entailments—it is important that we have the best possible understanding of the concept when developing instructions. Annotations or algorithms that strictly adhere to the proffered definition should not result in verdicts that all parties agree are misguided—which is exactly what could happen with a flawed definition.

This article is a review of the logical and philosophical issues involved in providing an adequate definition of textual entailment. We conduct a systematic analysis of logical accounts of textual entailment and discuss why each of them is ultimately unsatisfactory. We then examine both the advantages and the shortcomings of what has emerged as the standard definition of textual entailment, due to Ido Dagan and his collaborators (Dagan et al., 2005, 2009), and we articulate and defend a more adequate definition.

Logical Approaches to Textual Entailment

The terminology of "entailment," as well as at least some of the exemplary illustrations of textual entailment relationships, bring to mind the possibility that concepts such as *logical implication* or *deductive consequence* may be relevant to defining textual entailment. But, as many parties to this literature have observed, textual entailment cannot be adequately defined in such terms. Before turning to a more adequate, inferential approach, it is worth reviewing a variety of similarly named concepts in formal logic to see why textual entailment cannot be analyzed in terms of them.

Textual Entailment as Material Implication

It is not uncommon to express textual entailment in terms of natural language (indicative) conditionals: "if T then H." Elementary logic textbooks recommend representing natural language conditionals, at least for some purposes, as material conditionals (or "material implications"), which are compound propositions ($P \supset Q$) with a truth-functional compositional semantics: a material conditional is false when its antecedent is true and the consequent false, and it is true for the other three possible combinations of truth values (true/true, false/true, false/false).

Suppose that we defined textual entailment as material implication:

(D1) T textually entails H $=_{df}$ $T \supset H$

It may be that D1 gives an accurate necessary condition for textual entailment. But, as has been widely acknowledged, counterexamples to the sufficiency of material implication are easy to find. The definiens will be satisfied by any true H, regardless of T, and by any false T, regardless of H. So, for instance, the following will be textual entailments according to D1:

(T1) Albany is the capital of New York
(H1) Austin is the capital of Texas

(T2) Oswego is the capital of New York
(H2) Austin is the capital of Texas

(T3) Oswego is the capital of New York
(H3) El Paso is the capital of Texas

Each pair satisfies the definiens of D1, but would clearly not be considered cases of textual entailment; consequently, they reveal that the definiens ($T \supset H$) is not a sufficient condition for textual entailment.

Textual Entailment as Strict Implication

Material implication is not a sufficient condition for textual entailment. A narrower logical relationship, one that avoids the counterexamples above, is strict or logical implication (or in logic and philosophy, sometimes simply entailment). Intuitively, P strictly implies Q if it is (in an appropriately strong sense) impossible for it to both be the case that P is true and that Q is false.[3]

(D2) T textually entails H $=_{df}$ $\Box(T \supset H)$

The counterexamples to D1 considered above are not counterexamples to D2 because in each case it is possible for the antecedent to be true and the consequent false—if, for instance, the history of the United States had been other than it in fact was.

However, even though strict implication sets a higher bar than material implication, D2 has similar counterintuitive implications. For instance, D2 treats both of the following as textual entailments:

(T4) Oswego is the capital of New York and Oswego is not the capital of New York
(H4) El Paso is the capital of Texas

[2]Research in textual entailment grew rapidly between 2004 and the present, and there have been over 3,000 articles published on textual entailment and a series of influential conferences on automated identification of textual entailment. For just a sample, see Dagan and Glickman (2004), Dagan et al. (2005), Bar-Haim et al. (2006), Giampiccolo et al. (2007), Giampiccolo et al. (2008), Dagan et al. (2009), Androutsopoulos and Malakasiotis (2010), Sammons and Roth (2012), and Padó and Dagan (forthcoming). See Monz and de Rijke (2001) for an important precursor.

[3]Here we have in mind what philosophers would call "broadly logical impossibility" (Plantinga 1974, ch.1) or "impossibility *tout court*" (Kripke, 1980, p. 99), as distinguished from mere physical impossibility.

DOI: 10.1002/asi

(T5) Oswego is the capital of New York
(H5) El Paso is the capital of Texas or El Paso is not the capital of Texas

In each case it is impossible for the text to be true and the hypothesis false. In the first case this is because it is impossible for T4 to be true, from which it follows trivially that it is impossible for T4 to be true while H4 is false. In the second case this is because it is impossible for H5 to be false, from which it follows trivially that it is impossible for T5 to be true while H5 is false. Consequently, both fit the definition of strict implication and satisfy the definiens of D2. However, neither counts intuitively as a case of textual entailment.

A corollary is that automated theorem provers that strictly adhere to D2 are bound to deliver false predictions when dealing with contradictory texts and tautologous hypotheses. Because every tautology is strictly implied by everything, they will wrongly predict that every available tautologous text is a textual entailment of every text, and because every contradiction strictly implies everything, they will wrongly predict that every available contradictory text textually entails everything.

Textual Entailment as Relevant Implication

Both material implication and strict implication fail to provide a sufficient condition for textual entailment, largely due to the fact that both allow seemingly unrelated texts and hypotheses to count as textual entailments. The natural fix, then, is to require that T be, in some sense, relevant to H. One way of securing this is to make use of the logician's notion of relevant implication, where, very roughly, $p_1 \ldots p_n$ relevantly imply q iff (i) $p_1 \ldots p_n$ strictly imply q and (ii) the deduction of q from $p_1 \ldots p_n$ makes use of all of $p_1 \ldots p_n$. Thus, we get:

(D3) T textually entails H $=_{df}$ T relevantly implies H

This gives us the correct result in the case of T5/H5. The conclusion is a tautology that can just as well be derived from the empty set. Since the deduction of H5 does not make use of T5, T5 doesn't relevantly imply H5, and D3 gives the right result that this is not a textual entailment.

More obviously needs to be said about what it is to "make use of" a premise.[4] Fortunately, there is no need to fill in these details in order to see that the incorporation of relevance constraints is not enough to secure an intuitive account of textual entailment. Imagine an exemplary case of mathematical reasoning from some axioms $a_1 \ldots a_n$ to a highly complicated and nonobvious theorem t. Suppose further that the axioms are consistent and are all (in some intuitive sense) used in reasoning to the theorem. This should be a case of relevant implication if anything is. Thus, D3 will count this as a textual entailment:

(T6) $a_1 \ldots a_n$
(H6) t

But we can suppose that the mathematical deduction is quite abstruse—a major intellectual achievement, rather than the sort of natural inference associated with textual entailment. In fact, D3 will not only count such difficult mathematical deductions as textual entailments, it will count all relevant deductions of any kind, including those not yet achieved, and even those (if any) that are cognitively unachievable, as textual entailments. So relevant implication cannot be sufficient for textual entailment.

Textual Entailment as Doxastic Implication

One feature shared by all of the counterexamples to D1, D2, and D3 is that it is always possible for a reader to believe the one without believing the other. So perhaps the fix is to require, not that the text strictly implies the hypothesis, but rather that believing the text strictly implies believing the hypothesis. In other words, one text textually entails another just in case it is impossible for someone to believe the one without believing the other[5]:

(D4) T textually entails H $=_{df}$
$\Box(\forall x)[\text{believes}(x,T) \supset \text{believes}(x,H)]$

In short: textual entailment is doxastic implication.

There are at least two problems with this account. The first is that doxastic implication is far too demanding to serve as a constraint on textual entailment. Consider the following case:

(T7) There are Algerians in Paris
(H7) There are Algerians in France

This seems to be a textual entailment. And yet it is surely possible for someone to believe T7 without believing H7, perhaps because they are aware that there are Algerians in Paris but mistakenly believe that Paris is in Italy.

The second problem is that there are plausibly some propositions that are impossible to believe, for instance, *that something is green and nothing is green*. If that is right, then D4 predicts that the following is a textual entailment:

(T8) Something is green and nothing is green
(H8) El Paso is the capital of Texas

After all, it's impossible for someone to believe T8 without believing H8—as D4 requires—precisely because it's impossible for someone to believe T8 in the first place. But T8 does not textually entail H8; the two texts intuitively have nothing to do with one another. Thus, D4 faces the same problem as D1 and D2, in allowing unrelated texts and hypotheses to count as textual entailments. Doxastic implication is not sufficient for textual entailment.

[4]Cf. Mares (2012) for general discussion of relevance logic.

[5]See Renear (1988) for an account of how doxastic implication can be used to refine propositional relationships.

DOI: 10.1002/asi

As we'll see below, doxastic implication is not *entirely* off the mark. The correct analysis of textual entailment will make use of a related, but importantly different notion. Very roughly, it's not that believing the text *entails* believing the hypothesis, but rather that believing the text rationally commits one to believing the hypothesis.

Textual Entailment as Relevant Doxastic Implication

One might think that combining D3 and D4 will help:

(D5) T textually entails H $=_{df}$

i. $\Box(\forall x)[believes(x,T) \supset believes(x,H)]$ and

ii. T relevantly implies H

This yields the correct result that T6 does not textually entail H6. Condition (i) requires a certain level of cognitive proximity between T and H by requiring that it is impossible that T be believed and H not be believed. Accordingly, D5 is not susceptible to the counterexamples we raised against D3. D5 also correctly predicts that T8 does not textually entail H8: since T8 does not relevantly imply H8, condition (ii) is not met.

However, T7/H7 remains a problem. This is a genuine case of textual entailment, but, as indicated in in the previous section, it is possible to believe the one without believing the other. So condition (i) is not met; D5 specifies inaccurate necessary conditions for textual entailment.

Moreover, just as D3 and D4 can be combined into a single definition, their problems can likewise be compounded into a single counterexample. Once again, let $a_1 \ldots a_n$ be some simple axioms and let t be some complex theorem that is strictly implied by the axioms via some highly complicated deduction. Here is our counterexample:

(T9) $a_1 \ldots a_n$ & something is green and nothing is green
(H9) t & something is green and nothing is green

T9 does relevantly imply H9, since $a_1 \ldots a_n$ by hypothesis relevantly implies t. Moreover, it is impossible to believe T9 without believing H9, because it is impossible to believe T9's second conjunct and, thus, impossible to believe T9 as a whole. So both conditions of D5 are satisfied, and D5 thus predicts that T9 textually entails H9. But T9 does not textually entail H9, since $a_1 \ldots a_n$ don't textually entail t.

Inferential Approaches to Textual Entailment

What we saw in the preceding section is that it is a mistake to analyze textual entailment in terms of such logical notions as material implication, strict implication, doxastic implication, or relevant implication. An alternative approach is to analyze it in terms of inference. Dagan, Glickman, and Magnini (2005) developed an inferential definition for use in the PASCAL RTE challenges, and the following refined and widely-cited definition was advanced by Dagan, Dolan, Magnini, and Roth (2009):

(D6) T textually entails H $=_{df}$ typically, a human reading T would infer that H is most probably true

We begin by surveying some of the advantages of the inferential approach. We then turn to our main contribution: exposing the shortcomings of the definitions offered by Dagan and collaborators, articulating an improved definition that avoids these shortcomings, and defending our preferred definition against a range of objections.

Advantages of Inferential Analyses

Let us begin by examining three advantages of shifting to an inferential approach.

First, inferential analyses like D6 avoid several of the problems that arose for the logical analyses discussed above. Consider, for instance, one of our counterexamples to D1:

(T3) Oswego is the capital of New York
(H3) El Paso is the capital of Texas

In contrast to D1, D6 correctly predicts that T3 does not textually entail H3: people wouldn't typically infer that El Paso is probably the capital of Texas upon reading T3.

D7 also avoids the problem of impossible antecedents that plagued the analysis of textual entailment as strict implication:

(T4) Oswego is the capital of New York and Oswego is not the capital of New York
(H4) El Paso is the capital of Texas

In contrast to D2, D6 correctly predicts that T4 does not textually entail H4: people wouldn't typically infer that El Paso is probably the capital of Texas upon reading T4. Similar points apply to the problem of necessary consequents.

It likewise escapes the counterexamples to the doxastic implication analysis:

(T8) Something is green and nothing is green
(H8) El Paso is the capital of Texas

Unlike D4, D6 correctly predicts that this is not a textual entailment: people wouldn't typically infer that El Paso is probably the capital of Texas upon reading T8.

Second, inferential analyses like D6 are able to accommodate cases of textual entailment involving pragmatically generated implicatures. For instance, consider the following case:

(T10) Most of the passengers in the crash last year survived
(H10) Some of the passengers died in the crash

T10 does textually entail H10. But T10 does not strictly imply H10: T10 would still be true even if every passenger survived. So D2, D3, and D5 are all going to give the wrong result (as will D4, but for different reasons). D6, by contrast, gives the right result: a human reading T10 typically would infer that not all of the passengers survived. That's because

DOI: 10.1002/asi

"most" strongly suggests "not all." Reading T10, competent readers will naturally assume that not all the passengers survived, for in that case the author of the text would have said so and would not have made the weaker claim T10.[6] D6 is sensitive to the pragmatic norms that govern the inferential practices of ordinary readers.[7]

Third, inferential analyses like D6 are able to accommodate cases of textual entailment that rely on common background knowledge, including generic, geographic, and lexical knowledge.[8] For instance, T7/H7 should evidently be counted as a case of textual entailment:

(T7) There are Algerians in Paris
(H7) There are Algerians in France

And indeed, people reading T7 would typically infer that H7 is most probably true, since people typically know that Paris is in France.[9] So D6 makes the right prediction. But D2, D3, D4, and D5 all wrongly predict that this is not a textual entailment: T7 does not strictly imply H7 (Paris could come to be part of a different country) nor is it impossible to believe T7 without also believing H7.[10]

Refining the Inferential Analysis

The foregoing makes clear why we think that an inferential approach to textual entailment is on the right track. But D6, as stated, faces a series of problems that must be addressed by any satisfactory inferential analysis.

The first is the problem of *irrelevant trivialities*.

(T13) Lions are dangerous
(H13) I am reading something right now

Typical humans reading T13 would infer that (it is probably true that) they are reading something. So D6 implies that T13 textually entails H13. But it doesn't; T13 is about lions, not about you and what you are doing.

This problem can be ameliorated by "anchoring" the inference in T itself:

(D7) T textually entails H $=_{df}$ typically, a human reading T would infer *from T* that H is most probably true

In other words, for a text to entail a hypothesis, readers must be disposed to infer the hypothesis from the text itself. Readers of T13 do not infer that they are reading from T13 itself—the claim that lions are dangerous itself provides no support for the claim that they are reading something—but rather from the fact that they just read T13. Since D7 requires that the hypotheses be inferred from the text itself, it does not yield the result that T13 textually entails H13.

However, D7 still falls victim to a related problem. Consider the following text/hypothesis pair:

(T14) All your base are belong to us
(H14) The author of T14 is not a native English speaker

T14 does not textually entail H14. But H14 is something that typical readers would naturally infer from T14—specifically, from T14's bad grammar. So D7 wrongly predicts that T14 does textually entail H14.

To avoid this problem, we need to understand the envisaged inference as involving, not the strings T14 and H14, but rather as involving what is asserted by those strings: their contents, their meanings, the propositions they express. What we need, then, is D8:

(D8) T textually entails H $=_{df}$ typically, a human reading T would infer from the proposition expressed by T that the proposition expressed by H is most probably true

The proposition expressed by T14 (roughly, *that all of your bases belong to us*) does not entail or otherwise support the conclusion that its author is not a native English speaker. Thus, one would not typically infer what is asserted by an utterance of H14 from what is asserted by an utterance of T14. D8 correctly predicts that this is not a case of textual entailment. For ease of exposition, we will occasionally continue to speak of inferring hypotheses from texts, but this should be understood throughout as elliptical for talk of inferring what is expressed by one from what is expressed by the other.

The explicit reference to the proposition expressed also helps secure the right result for texts involving pronouns or

[6]Competent readers would also naturally assume that the fate of the remaining passengers is not unknown, for in that case one would expect the author to have qualified the statement: "we know that most passengers survived the crash; *indeed, perhaps all of them did*."

[7]See Zaenen et al. (2005) for further discussion of conversational and conventional implicature and its relation to textual entailment. Notice that the RTE5 guidelines take for granted that implicature is relevant to textual entailment: http://www.nist.gov/tac/2009/RTE/RTE5_Main_Guidelines.pdf. They say that T11 does *not* textually entail H11—

(T11) Yesterday 30 people were killed in a train accident near London.
(H11) 27 people died in a train accident.

—presumably because H11 is naturally read as saying that *only* 27 died. Yet T11 does require the truth of H11, since there cannot be 30 deaths without there being 27 deaths.

[8]See Bos and Markert (2005) for discussion of the varieties of background knowledge.

[9]We are assuming that this text is read in a context in which "Paris" is naturally read as referring to the European city, not, for example, to Paris, Texas. More on ambiguity in Background Knowledge, below.

[10]The same point can be made using an example from the RTE1 data set:

(T12) The Republic of Yemen is an Arab, Islamic, and independent sovereign state whose integrity is inviolable, and no part of which may be ceded.
(H12) The national language of Yemen is Arabic.

RTE1 designates this as a genuine textual entailment. But T12 does not logically entail H12; it is logically possible for Yemen to be a sovereign Arab state and yet have no national language. So D2 predicts that T12 does not textually entail H12. D6, by contrast, does predict that T12 textually entails H12: people would typically infer that Arabic is most probably the national language of Yemen from the information in T12 together with their background knowledge about national languages (for example, that countries typically have one). Cf. Dagan et al. (2009: v).

DOI: 10.1002/asi

ambiguous terms. For instance, does the text "Angela Merkel is the Chancellor of Germany" textually entail the text "She is a German politician"? That depends on whether the latter text expresses a proposition about Angela Merkel. If it does, and if the context makes this apparent to a typical reader, then this is a case of textual entailment; if not, then it isn't a case of textual entailment. This is just to say that D8 is poised to make the right predictions in such cases. Although it clearly leaves open the fraught philosophical question of what determines the referent of a pronoun or demonstrative in a given context, as well as the difficult practical question of how to build an algorithm that can identify the intended referent of a pronoun or the intended sense of an ambiguous term.[11]

It should be noted that Dagan et al. (2009) appear to recognize the need for these first two modifications. Elsewhere, Dagan et al. (2005) provide an importantly different characterization of textual entailment:

(D9) T textually entails H $=_{df}$ the meaning of *H* can be inferred from the meaning of *T*, as would typically be interpreted by people[12]

This incorporates both the needed anchoring of the inference in T itself and the needed reference to the propositions expressed by T and H.

However, while avoiding the problems we have raised for D6, D9 has problems of its own. The most serious problem stems from the fact that, unlike D6 which is framed in terms of what *would* be inferred from a text, D9 is framing in terms of what *can* be inferred from a text. Accordingly, the counterexample to D3—in which the hypothesis can be inferred from the text but only by some extraordinarily complicated line of reasoning—turns out to be a counterexample to D9 as well. Likewise, the mere fact that it is possible for some confused individual to infer from T15 that H15 is likely true would (according to D9) suffice for textual entailment:

(T15) Tom is from Paris
(H15) Tom is from Italy

But clearly there is no textual entailment here. D8 does not make these unwanted predictions, since it requires that typical readers *would* draw the inference, not just that someone *could*. D8 is an improvement on D6 and D9, incorporating the desirable features of both and eliminating some of the undesirable features.

We have seen that D8 avoids the problems posed for other inferential approaches examined thus far. But D8 is problematic as well, since it faces a problem of overlooked entailments:

(T16) Charles was enjoying a relaxing bath
(H16) Charles does not have ablutophobia, the extreme fear of bathing

T16 does evidently textually entail H16. But it would not typically occur to someone to draw any inferences about whether Charles is an ablutophobe. The fear of bathing typically wouldn't even cross their mind when reading T16; most people haven't even heard of ablutophobia. Put another way, inferring that Charles most probably does not have ablutophobia is not something that a human would typically do upon reading T16. So D8 wrongly predicts that T16 does not textually entail H16.[13]

The problem can be fixed by requiring only that one be *justified* in making the inference. To say that someone is justified in inferring p from q is, roughly, to say that it is reasonable for her/him to make the inference, given what s/he knows about p and q and his/her other background knowledge.[14] Crucially, it can be true that one would be justified in making an inference even when one hasn't in fact made the inference.[15] Here then is the needed revision:

(D10) T textually entails H $=_{df}$ typically, a human reading T would be justified in inferring from the proposition expressed by T that the proposition expressed by H is most probably true

One would be justified in inferring H16 from T16, even though it is highly unlikely that one would actually make the inference. So D10 delivers the right result, that T16 does textually entail H16.[16]

There is one further problem, which plagues all of the inferential analyses surveyed thus far, including D10. This is the problem of uninferrable likelihoods:

(T17) John entered a million-ticket raffle
(H17) John lost the raffle

T17 does not textually entail H17: T17 leaves it entirely open who won the raffle. But one who reads T17 would typically infer that H17 is most probably true, and this inference would indeed be justified. So D10 wrongly predicts that H17 is textually entailed by T17.

This brings us to what we take to be the correct analysis (*modulo* one final adjustment in Background Knowledge, below):

[11]See Lewis (1979), Kaplan (1989), and Braun (2015, §1.4) on philosophical challenges, and see Mirkin et al. (2010) on challenges for implementation.

[12]Actually, they present this only as a sufficient condition for textual entailment, and so may not intend for it to be a definition.

[13]We choose an example involving such an obscure phobia in part to head off objections to the effect that one might sometimes nonconsciously infer a hypothesis from a text. But since people typically do not even have the concept *ablutophobia*, it can't be that people typically infer anything about ablutophobia, even nonconsciously.

[14]See Feldman (2002: ch. 4–5) for a useful introduction to the notion of justification.

[15]Cf. Feldman (2002: 46) on the difference between having a belief that is justified and being justified in forming that belief.

[16]Alternatively, one might modify D6 (or D7 or D8) to say "T textually entails H $=_{df}$ typically, a human reading T *and H* would infer. . .." It may be that this is what Dagan and collaborators had intended for D6 to say. We are grateful to an anonymous referee for this observation.

DOI: 10.1002/asi

(D11) T textually entails H $=_{df}$ typically, a human reading T would be justified in inferring the proposition expressed by H from the proposition expressed by T

All that we have done is drop the reference to probability, so that now what one must be justified in inferring is that H is true, not merely that it is highly likely to be true. And while one would be justified in inferring that H17 is probably true, one would not be justified in inferring H17 straight-out—someone wins the raffle, after all, and there is no good reason to think it isn't John. So D11 rightly predicts that T17 does not textually entail H17.

Clarifications and Complications

In this section we test our proposed definition of textual entailment against some challenging text/hypothesis pairs. Through these tests, we also uncover some interesting formal properties of the textual entailment relation. In particular, we see that textual entailment is context-sensitive, intransitive, and nonmonotonic.

Inference and Belief

Here is an important clarification about the relationship between inference and belief. Normally, when one infers some claim B, one thereby comes to believe B. But not always.

There are at least two sorts of cases in which one infers one claim from another without believing it. The first sort of case involves reasoning from a claim that one knows to be false to another that one knows to be false. For instance:

(T18) The moon is made of green cheese
(H18) The moon is made of cheese

One can justifiably infer H18 from T18. But that doesn't mean that one is thereby justified in *believing* H18. Inferring does not entail believing. We do this routinely in *reductio ad absurdum* reasoning: we begin by assuming (for the sake of argument) something we believe to be false and then demonstrate its falsity by inferring other claims from it that are indisputably false. If all goes well, one is justified in inferring each step from the preceding step. But obviously one does not believe the absurdity that one ultimately infers.

The second sort of case involves rejecting a claim upon realizing what follows from it. For example:

(T19) Killing is always wrong
(H19) Killing a life-threatening tapeworm is wrong

One can imagine an overzealous pacifist initially embracing T19. But the pacifist will see that T19 is false as soon as she recognizes that it entails the absurd claim H19. In other words, she infers H19 from T19 and thereby comes to see that T19 is false. She is justified in drawing that inference, since H19 does follow from T19. But she would not be justified in believing H19 upon reading T19. H19 is absurd, and once she realizes that H19 can justifiably be inferred from T19, the only rational response is to reject T19 (and perhaps retreat to a less stringent prohibition on killing).

Thus, D11 should be sharply distinguished from a nearby analysis D12:

(D12) T textually entails H $=_{df}$ typically, a human reading T would be justified in believing the proposition expressed by H on the basis of the proposition expressed by T

Typical humans know that T18 is false, and so wouldn't be justified in believing H18 on the basis of T18. And typical humans would know that something has gone wrong if H19 follows from something they believe, so they wouldn't be justified in believing H19 on the basis of T19. D12 therefore wrongly predicts that there is no textual entailment between T18 and H18 or between T19 and H19. D11, by contrast, gets the right result: one would be justified in inferring H18 and H19 from T18 and T19 (respectively), even though one would not be justified in believing H18 or H19.

D11 should be glossed not in terms of D12 but rather in terms of D13:

(D13) T textually entails H $=_{df}$ typically, a human is justified in reasoning from the proposition expressed by T to the proposition expressed by H

One can reason from one proposition to another without thereby believing either proposition. And that is how inferring is to be understood in D11.

Background Knowledge

As we saw in Advantages of Inferential Analyses, above, when assessing whether one text is textually entailed by another, we need to take into account background knowledge that readers are likely to have. For instance, we want to count the following as a case of textual entailment:

(T7) There are Algerians in Paris
(H7) There are Algerians in France

And D11, like D6, gets the right result in this case: people reading T7 are typically justified in inferring H7, since people typically know that Paris is in France.

But matters get complicated when we consider the extent to which background knowledge should figure into our assessments of textual entailment. Consider, for instance, the following case of expert background knowledge:

(T20) Ferrous sulfate heptahydrate is green
(H20) $FeSO_4 \cdot 7H_2O$ is green

There is reason to think that this is a case of textual entailment. We would certainly want information extraction applications for research databases to deliver results about ferrous sulfate heptahydrate when queried about $FeSO_4 \cdot 7H_2O$. And

DOI: 10.1002/asi

yet, D11 seems to give the opposite result. It is not typically the case that humans—most of whom have no idea that these are names for the same chemical compound—would be justified in inferring H20 from T20.

Similar issues arise in connection with local background knowledge.

(T21) Phil Rudd is from Tauranga
(H21) Phil Rudd is from New Zealand

New Zealanders will typically know that Tauranga is in New Zealand; it's one of New Zealand's largest cities. But most humans have never heard of Tauranga and, so, won't be justified in inferring H21 from T21. Yet there is reason to treat this as a case of textual entailment. We would, for instance, want a search application for the *New Zealand Herald's* database to deliver T21 in response to the query "Is Phil Rudd from New Zealand?" So D11 seemingly gives the wrong result.

But, in defense of D11, imagine somebody without any chemistry background who needed, for one reason or another, to know whether or not $FeSO_4{\cdot}7H_2O$ is green. The application would not be returning useful information if it returned a text containing "Ferrous sulfate heptahydrate is green." And similarly with T21/H21. If, for whatever reason, some non-New Zealander wanted to know whether or not Phil Rudd was from New Zealand, the information retrieval (IR) application would not be returning the desired information if it returned the text "Phil Rudd is from Tauranga."

So there are good reasons to think that T20/H20 and T21/H21 should count as textual entailments, and there are good reasons to think they should not. The reasons for or against seem to depend on what sort of end user we have in mind. In the case of T20 and H20, our intuitions about textual entailment seem to depend on whether we have chemists in mind; in the case of T21/H21, the intuition depends on whether or not we have New Zealanders in mind

A way to accommodate both intuitions is to recognize that textual entailment is relative to a variable group of end users. More precisely,

(D11*) T textually entails H relative to group G $=_{df}$ typically, a member of G reading T would be justified in inferring the proposition expressed by H from the proposition expressed by T

D11* correctly predicts that T20 entails H20 if we have in mind an application for chemists, but that T20 does not entail H20 if we have in mind end users with no training in chemistry. (*Mutatis mutandis* for T21/H21.)

Relativizing the definition of textual entailment to G is a change for the better. It provides a mechanism for capturing pairs like T20/H20 and T21/H21 as textual entailments relative to a group G_{Chem} of trained chemists, where a group G_{Ann} of annotators who are not trained chemists would have recognized no entailment. The above definition D11 can itself be construed as an instance of D11*, defining textual entailment for humans in general.

The sensitivity of textual entailment to a group G does not necessarily force a deep revision to the RTE project. The data sets against which textual entailment software is tested are determined by annotators whose backgrounds can vary. As we saw above, at least some background facts in these domains need to be available for any textual entailment software. For example, an automated theorem prover would have background knowledge encoded as axioms for use in deductions. In one specific example of a theorem prover (Bos & Markert, 2005), geographic background knowledge is taken directly from the CIA factbook. Of course, as the examples above illustrate, there is a danger in making too much background knowledge accessible. But accounting for this, and accommodating the sensitivity of textual entailment to a given group of end users more generally, simply requires varying which background facts are accessible to the application.

(For ease of exposition, we will ignore these complexities for the remainder of the article, and continue to focus on the oversimplified definition D11 rather than the more adequate D11*.)

Inferential Distance

Recall our example from Textual Entailment as Relevant Implication, above, of reasoning from some axioms $a_1 \ldots a_n$ to a highly complicated and nonobvious theorem t, provable in no less than 100 deductive steps of inference. And consider the following text/hypothesis pair:

(T6) $a_1 \ldots a_n$
(H6) t

On the face of it, this is not a case of textual entailment. Textual entailment is meant to be sensitive to practical considerations. Suppose, for example, that a user of a question answering application wants to know whether H6 is arithmetically true, and that application returns T6. T6 is not a useful output for the user, because it doesn't answer her question.

To help see this, consider the fundamental theorem of arithmetic: that every nonprime integer greater than 1 is a unique product of primes. The proof from the ZFC axioms is incredibly complicated but can be laid out in such a (tedious) way that each step of the proof textually entails the preceding step. Now imagine a user's disappointment if, in response to the query "is every nonprime integer greater than 1 a unique product of prime numbers?" an Internet search engine led her to a Wikipedia page that simply lists the ZFC axioms. She would be disappointed because this is (for any normal user) an utterly useless output. The ZFC axioms may logically entail an affirmative answer to the query, but they do not *textually* entail one.

This is a case in which Dagan et al.'s definition D6 straightforwardly gets the right result. It is not true that,

DOI: 10.1002/asi

typically, a human reading T would infer that H is most probably true, so D6 correctly predicts that T6 does not textually entail H6. But one might object that our preferred definition D11 makes the wrong prediction. After all, there is a justification for inferring H6 from T6, namely, the aforementioned proof. So, the idea goes, D11 wrongly predicts that T6 textually entails H6.

But this reasoning rests on a confusion between there being a justification (or reasons) for inferring something and one having a justification (or reasons) for inferring it. To help see this, notice that there can be a justification for *believing* something even though one does not oneself have that justification. For instance, if the news channels have all reported that Smith was found guilty, but one hasn't been watching the news, then there is a justification for believing that Smith was found guilty but one does not have that justification. To have a justification, it's not enough that there simply be reasons; one must also be aware of those reasons. And this is how justified inference is to be understood in D11: one is justified in inferring H6 from T6 only if one has a justification for inferring H6 from T6. So understood, D11 makes the right prediction: typically a human is not justified in inferring H6 from T6, and so T6 does not textually entail H6.

These sorts of examples also bring out an interesting and perhaps surprising point about textual entailment, namely, that it is not transitive (at least not in the strict mathematical sense).[17] To see this, suppose that each step in the proof follows trivially from the previous step alone. Each step (understood as a fragment of text) will then textually entail the subsequent step, and the penultimate step will textually entail H6. If textual entailment were transitive, then it would follow that T6 textually entails H6. But it doesn't. So textual entailment is not transitive.

Monotonicity

We just saw that textual entailment is not transitive. Now let us turn to some examples that purport to show that textual entailment is nonmonotonic. That is, merely adding information to a text can result in *removing* some of its textual entailments.

Consider T22, a passage from *The Giving Tree*:

(T22) And so the boy cut down her trunk. And made a boat and sailed away. And the tree was happy.
(H22) The tree was happy.

T22 plainly does textually entail H22. And D11 bears this out: humans would typically be justified in inferring H22 from T22. Now consider the continuation of the passage:

(T22*) And so the boy cut down her trunk. And made a boat and sailed away. And the tree was happy. But not really.

People reading T22* would not be justified in inferring H22. Rather, they would know, having read the final sentence, that the author is being facetious in the third sentence. Thus, adding information to T22 without subtraction turns the entailment into a nonentailment.

Another example:

(T23) Bill is so responsible.
(T23*) Bill was late to the meeting because, once again, he forgot to set his alarm. Bill is so responsible.
(H23) Bill is responsible.

T23 plainly textually entails H23, and T23* plausibly doesn't. Readers will typically know that the second sentence in T23* is sarcastic, and they won't take the author to be saying that Bill is in fact responsible.[18]

This reinforces a point already made in Logical Approaches to Textual Entailment in connection with logical analysis. Algorithms for testing for textual entailment that blindly treat any theorem as a textual entailment are prone to false positives—in this case, because they are blind to sarcasm and other pragmatic effects. Just as one needs to limit what background information is available (see Background Knowledge, above), one needs to limit which logical implications are to count as textual entailments.

We have seen that D11 predicts failures of monotonicity in cases involving inconsistent texts. However, there are other cases involving inconsistent texts where we intuitively do get textual entailments. Consider, for example, a passage from a verbatim transcript of a deposition of a witness to a crime:

(T24) The suspect and I were at a friend's house at 7:30 pm on April 2, 2015 [. . .] We were at the gym at 7:30 pm on April 2, 2015.
(H24) The suspect and I were at a friend's house at 7:30 pm on April 2, 2015.

Intuitively, T24 does textually entail H24; we would, for instance, expect an information extraction application to treat H24 as a textual entailment of T24. And, indeed, D11 correctly predicts that this is a case of textual entailment: a human reading T24 is justified in inferring the proposition expressed by H24 from the proposition expressed by T24. (This is not to say that one would be justified in *believing* H24 upon reading the inconsistent testimony in T24; as we saw in Inference and Belief [above], being justified in inferring something doesn't require being justified in believing it.) The difference between T23 and T24 that allows for this differential treatment is that the occurrence of H23 in T23* is naturally taken to be nonliteral speech (viz., sarcasm), whereas the occurrence of H24 in T24 is naturally taken to be entirely literal (though perhaps a mistake or a lie). It is a

[17] *Pace* Berant et al. (2012).

[18] Here we are assuming that a sarcastic utterance of "Bill is so responsible" expresses the proposition that Bill is irresponsible, and thus expresses a different proposition from H23.

DOI: 10.1002/asi

virtue of D11 that it is able to mark this distinction between different kinds of inconsistent texts.

Conclusion

It is crucial that research on textual entailment be underwritten by the best possible understanding of textual entailment itself. We have argued that, while the inferential approach to defining textual has clear advantages over logical approaches, the (now) standard definition proffered by Dagan and collaborators is in need of refinement. We have articulated and defended a refined definition, still in the spirit of Dagan et al.'s, according to which a text T textually entails a hypothesis H relative to a group of end users G just in case, typically, a member of G reading T would be justified in inferring the proposition expressed by H from the proposition expressed by T. We have shown how our definition improves upon existing definitions, and we have defended it against a range of objections. Finally, we argued that textual entailment is context-sensitive, nontransitive, and nonmonotonic. This clarification of the notion of textual entailment may be considered, more generally, as an exercise in the conceptual foundations of information science.

Acknowledgments

We acknowledge the contributions by David Dubin, Catherine Blake, and Henry A. Gabb, all at the School of Information Sciences, University of Illinois at Urbana-Champaign, as well as two anonymous referees.

References

Androutsopoulos, I. & Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. Artificial Intelligence Research, 38, 135–187.

Bar-Haim, R. et al. (2006). The second PASCAL Recognising Textual Entailment Challenge. Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment. Venice, Italy. Retrieved June 19, 2015 from http://u.cs.biu.ac.il/~nlp/RTE2/Proceedings/01.pdf

Berant J., Dagan, I. & Goldberger, J. (2012). Learning entailment relations by global graph structure optimization. Computational Linguistics, 38(1), 73–111.

Blake, C. (2011). Text mining. Annual Review of Information Science & Technology, 45, 123–155.

Bos, J. & Markert, K. (2005). Recognizing textual entailment with logical inference. Proceedings of the Conference on Human Language Technology and Emirical Methods in Natural Language Processing, 628–635.

Braun, D. (2015). Indexicals. The Stanford Encyclopedia of Philosophy (Summer 2017 Edition), E.N. Zalta (ed.), https://plato.stanford.edu/archives/sum2017/entries/indexicals/

Dagan, I., Dolan, B., Magnini, B. & Roth, D. (2009). Recognizing textual entailment: Rational, evaluation and approaches. Natural Language Engineering, 15(4), i–xvii.

Dagan, I. & Glickman, O. (2004). Probabilistic textual entailment: generic applied modeling of language variability. PASCAL Workshop on Learning Methods for Text Understanding and Mining. Grenoble, France. Retrieved from http://u.cs.biu.ac.il/~dagan/publications/ProbabilisticTE_fv07.pdf

Dagan, I., Glickman, O. & Magnini, B. (2005). The PASCAL Recognizing Textual Entailment Challenge. MLWC'05 Proceedings of the First International Conference on Machine Learning Challenges: Evaluating Predictive Uncertainty Visual Object Classification, and Recognizing Textual Entailment (Southampton, UK, 11–13 April 2005), 177–190.

Feldman, R. (2002). Epistemology. New York: Pearson.

Giampiccolo, D. et al. (2008). The Fourth PASCAL Recognizing Textual Entailment Challenge. Proceedings of the TAC 2008. NIST, Gaithersburg, MD. Retrieved from http://www.nist.gov/tac/publications/2008/additional.papers/RTE-4_overview.proceedings.pdf

Giampiccolo, D., Magnini, B., Dagan, I. & Dolan, B. (2007). The Third PASCAL Recognizing Textual Entailment Challenge. Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Prague, Czech Republic. Retrieved from http://dl.acm.org/ft_gateway.cfm?id=1654538

Gómez-Adorno, H., Pinto, D. & Vilariño, D. (2013). A question answering system for reading comprehension tests. Lecture Notes in Computer Science, 7914, 354–363.

Harabagiu, S., Hickl, A. & Lacatusu, F. (2007). Satisfying information needs with multi-document summaries. Information Processing and Management, 43(6), 1619–1642.

Kaplan, D. (1989). Demonstratives. In Themes from Kaplan, J. Perry & H. Wettstein (eds.), pp. 481–563. Oxford: Oxford University Press.

Kolterman, L., Dagan, I., Magnini, B. & Bentivogli, L. (2015). Textual entailment graphs. Natural Language Engineering, 21(5), 699–724.

Kripke, S. (1980). Naming and Necessity. Cambridge, MA: Harvard University Press.

Lewis, D. (1979). Scorekeeping in the Language Game, Journal of Philosophical Logic, 8, 339–359.

Magnini, B., Dagan, I., Neumann, G. & Padó, S. (2014). Entailment graphs for text analytics in the excitement project. Lecture Notes in Computer Science, 8655, 11–18.

Mares, E. (2012). Relevance Logic. The Stanford Encyclopedia of Philosophy (Spring 2014 Edition), E.N. Zalta (ed.). Retrieved from https://plato.stanford.edu/archives/spr2014/entries/logic-relevance/

Mirkin, S., Dagan, I. & Padó, S. (2010). Assessing the role of discourse references in entailment inference. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 1210–1219.

Mohler, M., Tomlinson, M. & Bracewell, D. (2013). Applying textual entailment to the interpretation of metaphor. In Proceedings of the 7th International Conference on Semantic Computing. Irvine, CA.

Monz, C. & de Rijke, M. (2001). Light-weight entailment checking on inference in computer semantics. Proceedings of the Conference on Inference in Computational Semantics, 59–72.

Padó, S., Cer, D., Galley, M., Jurafsky, D. & Manning, C.D. (2009). Measuring machine translation quality as semantic equivalence: A metric based on entailment features. Machine Translation, 23(2–3), 181–193.

Padó, S. & Dagan, I. (forthcoming). Textual entailment. In R. Mitkov (ed.) Oxford Handbook of Computational Linguisitics. Oxford, UK: Oxford University Press.

Plantinga, A. (1974). The nature of necessity. Oxford, UK: Oxford University Press.

Renear, A H. (1988). The Paradoxes of Doxastic Implication: An Essay on the Logic of Belief (Doctoral dissertation, Brown University).

Renear, A.H. & Palmer, C.L. (2009). Strategic reading, ontologies, and the future of scientific publishing. Science, 325(5942), 828–832.

Sammons, M., Vydiswaran, V. & Roth, D. (2012). Recognizing textual entailment. In D.M. Bikel & I. Zitouni (eds.) Multilingual Natural Language Applications: From Theory to Practice. Upper Saddle River, NJ: IBM Press, pp. 209–258.

Udayakumar, R., Khanaa, V. & Thooyamani, K.P. (2014). Fresh information retrieval using P2P web search. Middle-East Journal of Scientific Research, 20(12), 1904–1907.

Zaenen, A., Karttunen, L. & Crouch, R. (2005). Local Textual Inference: can it be defined or circumscribed? In Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment. Ann Arbor, MI. Retrieved from http://dl.acm.org/citation.cfm?id=1631868