Toward Annotating Commonsense Inferences in Text: INCOMPLETE DRAFT

Ernest Davis

Abstract

The objective of the TACIT (Toward Annotating Commonsense Inferences in Text) project is to identify all or most of the commonsense inferences needed to understand a small collection of short narrative texts; to characterize those inferences in terms of features in different dimensions; and to characterize the commonsense knowledge that underlies those inferences. The primary purpose of this analysis is to help map out the space of commonsense knowledge as a guide for research in knowledge representation. Secondarily, the corpus could be used to evaluate progress in automated commonsense reasoning and in the integration of commonsense reasoning into automated natural language interpretation. To date, we have developed a framework for the annotation and a standard representation in XML; and we have analyzed T short texts with a total of S sentences and characterized Q commonsense inferences in those texts.

1 Introduction

The objective of the TACIT (Toward Annotating Commonsense Inferences in Text) project is to identify all or most of the commonsense inferences needed to understand a small collection of short narrative texts; to characterize those inferences in terms of features in different dimensions; and to characterize the commonsense knowledge that underlies those inferences. We have developed and put on the web an XML-based corpus of T short texts, extracted from news stories, with a total of S sentences, in which we have identified a total of Q inferences.

The primary goal of the project is to help map out the space of commonsense knowledge, as a guide for research in knowledge representation. Our intent is that, by careful and exhaustive analysis of the commonsense reasoning used in real-world texts, we can get a clear and realistic idea of what kinds of knowledge and reasoning would actually be needed for an important category of real-world task.

In general, the study of automated commonsense reasoning has often suffered from the lamppost problem. Formalist approaches to commonsense reasoning have tended to focus on problems that lend themselves to elegant formalization (this has been particularly acute in such areas as spatial reasoning (Cohn & Renz 2007) and nonmonotonic reasoning (Brewka, Niemelli, & Truszczynsky 2007)); web-based approaches tend to focus on information that is easily extracted such as taxonomic relations (Banko et al. 2007); and crowd-sourcing approaches (Havasi et al. 2010) tend to end up collecting information that, for one reason or another, is salient to naïve contributors rather than of fundamental importance in reasoning. There is, of course, something to be said for lamppost-centered research — it is often better to make some progress on peripheral problems where progress can be made than to beat your head against problems where nothing can be accomplished, however central these are — but it is also important to keep in mind the larger view. Research enterprises such as RTE and SemEval also address the issues of reasoning in text interpretation, but, as we will discuss in detail in section 7, the scope and objectives of these projects is significantly different.

As a secondary goal, the TACIT corpus could potentially serve as the basis for evaluating, either knowledge bases for commonsense knowledge, such as CYC (Lenat, Prakash, & Shepherd 1985) or ConceptNet (Havasi et al. 2010), or natural language understanding systems that attempt to incorporate commonsense reasoning. Partly with this end in mind, and partly as an aid to systematizing the inferences, we have included with each inference one or more multiple choice questions, designed to test whether a reader has correctly performed the inference.

It is not intended that this corpus should serve as a training corpus for any plausible machine learning program. Not only is it too small, but, much more seriously, the features are not nearly sufficiently well-defined or systematic, as discussed in section 5.

The objective, as stated above, is to characterize the commonsense inferences involved in natural language understanding, as distinguished from the other aspects of natural language understanding. Therefore, in developing annotations we adopt the following standpoint. Imagine that all aspects of natural language understanding *except* the integration of commonsense reasoning had been solved. We have a text reading program with a complete lexicon (which, of course, has been achieved in reality); a perfect understanding of syntax (largely achieved); and a perfect understanding of semantics and its relation to syntax (far from completely

achieved, but for argument's sake). What the hypothetical program lacks is any understanding of the real world or of the purposes of communication. It can use simple heuristics for selecting the correct interpretation of lexically ambiguous words, such as "choose the most common meaning" and similarly simple heuristics for other forms of ambiguity, such as "resolve references to the nearest possible previous referent." But it cannot resolve any ambiguity on the basis of how likely it is that the interpretation would be true, or how likely it is that someone would say it. The task we are addressing in TACIT is, given a program of this description, what *additional* knowledge is needed to interpret the text in hand? Of course, this characterization has to be taken with plenty of salt, since it posits a sharp distinction between linguistic knowledge and world knowledge that is imaginary.

Having identified an inference that needs to be made, the annotator tries to answer the following questions:
1. What background knowledge does the human reader use in order to resolve the gap? We are primarily interested here in knowledge of facts about the external world and discourse conventions, not in facts about the language (e.g. the meanings of idioms).
2. What is the domain of the background knowledge in (1)?
3. Why is it important to carry out this inference? What role does the inference play in the process of understanding? We try to ground this answer as far as possible in linguistic theory and theory of narrative; hence we call it the "linguistic significance".
4. What is the logical structure of the fact being inferred?

The claim that we are characterizing *all* the necessary inferences and, still more the claim that we are characterizing *all* the necessary background knowledge, have to be taken even more loosely. The facts listed as "background knowledge" in our corpus of annotations do not come close to be a full account of what in fact be needed by a knowledge-based reasoner to carry out the inferences; there are all kinds of underlying knowledge, particularly in fundamental domains such as space, time, and naïve psychology that is omitted. For instance, we have nowhere included frame axioms, which in practice are always critical. Additionally, a real knowledge-based reasoner for this purpose would face the problem of integrating rules of plausible inference that point in different directions; we do not address that problem at all. Rather, our list of "background facts" contains a few facts closely associated with the inference that are particularly salient and unusual.

2 Examples

The project is best explained with examples. The first text in our corpus is the following:

> On a mundane morning in late summer in Paris, the impossible happened. The Mona Lisa vanished. On Sunday evening, August 20, 1911, Leonardo da Vinci's best-known painting was hanging in her usual place on the wall of the Salon Carré between Correggio's Mystical Marriage and Titian's Allegory of Alfonso d'Avalos. On Tuesday morning, when the Louvre reopened to the public, she was gone. Within hours of the discovery of the empty frame, stashed behind a radiator, the story broke in an extra edition of Le Temps, the leading morning newspaper. Incredulous reporters from local papers and international news services converged on the museum. Georges Benedité, the acting director, and his curators were speculating freely to the press.

We have identified thirty-four inferences involved in understanding this text that require commonsense knowledge. Our description of inferences #1, 3, and 4 is as follows (inference 2 is less interesting).
Inference 1: In "the impossible happened", "impossible" is hyperbolic, not literal. What is meant is "a very improbable event".
Specific text being explicated: "the impossible happened".
Background: An impossible event cannot happen.
Category of Inference: (PropertyOf=Unlikely; Event="the impossible";)
Domain: Necessity and possibility.
Linguistic Significance: Interpret non-literal text. **Question:** How likely did the event under discussion seem before it occured? **Right answer:** Quite unlikely. **Wrong answers:** (a) Impossible. (b) Likely. (c) Certain.
Question: How likely is it now that the event under discussion occured? **Right answer:** Certain. **Wrong answers:** (a) Likely. (b) Quite unlikely. (c) Impossible.

****

Inference 3: In "The Mona Lisa vanished", "vanished" is metaphorical, not literal. What is meant is "The Mona Lisa became absent from its proper place". **Specific text being explicated:** "The Mona Lisa vanished".
Background: Physical objects rarely literally vanish.
Category of Inference: (Existence; Event=Mona Lisa became absent;)
Domain: Spatial and physical knowledge.
Linguistic Significance: Interpret non-literal text.
Question: What actually happened to the Mona Lisa?
Right answer: The Mona Lisa unexpectedly became missing from its usual place. **Wrong answer:** The Mona Lisa became invisible.

****

Inference 4: The event of the Mona Lisa leaving its place and the event judged to be impossible in sentence 1 are the same event.
Specific text being explicated: "...the impossible happened. The Mona Lisa vanished".
Background:
1. It is important that valuable objects remain where they are supposed to be, and great efforts are made to ensure that they do so. Therefore, it is considered highly improbable that a valuable object will leave its place, other than under the supervision of the authorities responsible for it.
2. A painting in a museum is a valuable object.
3. Paintings in a museum are under the supervision of the museum administrators

Compare: "... the impossible happened. A bar of soap had vanished from the men's bathroom at the Louvre."

Category of Inference: (Identical; Event="the impossible"; Event=Mona Lisa vanished;)
Domain: Organizations. Property.
Linguistic Significance: Coreference resolution.
Additional Linguistic Clues: The metaphorical "vanished" fits with the hyperbolic "impossible"; it would be literally impossible for the Mona Lisa to literally vanish.
Question: What is the connection between "the impossible happened" and "The Mona Lisa vanished"? **Right Answer:** The Mona Lisa vanishing is the nearly impossible event that happened. **Wrong Answer:** First the impossible happened, then later the Mona Lisa vanished.
Question: Why was it considered nearly impossible for the Mona Lisa to be missing? **Right answer:** Because museums try hard to make sure that their valuable artworks are always in the proper place. **Wrong answer:** Because paintings do not usually vanish.

Hopefully these are reasonably self-explanatory. The first two inferences that need to be made are that "impossible" and "vanished" are figurative, not literal. The third inference is that the phrases "the impossible happened" and "the Mona Lisa vanished" refer to the same event. The linguistic significance of the first two is to interpret non-literal text; the linguistic significance in the third is coreference resolution (determining that two entities mentioned in the text are the same). The first requires general knowledge that impossible things cannot in fact happen; the domain of this fact comes is the general theory of necessity and possibility. The second requires the more specific knowledge that physical objects rarely literally vanish; this comes from a physical theory. We categorize the conclusion in the first as the inference that the event (whatever it is) denoted as "the impossible" has the property of being unlikely. We categorize the conclusion in the second as an inference that the event of Mona Lisa becoming absent occurred (existed). The format we use for these e.g. "(PropertyOf = Unlikely ; Event = "the impossible" ;)" is explained in the next section.

The third inference is substantially more complex. Having interpreted "the impossible happened as "a very unlikely event occurred" and having interpreted "the Mona Lisa vanished" as "the Mona Lisa became absent from its usual place in an unexpected way", the reader must now connect the two. This involves understanding why it is that the unexpected absence of the Mona Lisa would be considered so extremely unlikely; as the sentence introduced as a point of comparison illustrates, if a bar of soap unexpectedly became absent, one would hardly describe that at "the impossible happening" except as a joke. This understanding thus depends on an understanding of the value of famous paintings and the care that is taken to make sure that their whereabouts are always known to the responsible authorities. We have formulated this knowledge in the background facts numbers 1-3; obviously, the individuation as separate facts is somewhat arbitrary. We characterize the home domain of these facts as partly in the theory of property and partly in the theory of organizations.

The linguistic significance of this inference comes under the category of coreference resolution; we need to determine that two separate (and quite different) phrase in the text in fact refer to the same entity (the Mona Lisa becoming absent). We categorize the type of inference as the statement that the two events "the impossible" and the Mona Lisa vanishing are in fact identical.

We note further, under "Additional linguistic clues" that this interpretation receives further support from the fact that the writer is continuing the same figure of speech; having said that the event is impossible, he describes it in terms that are, in fact, impossible.

With each inference, we include one or more multiple-choice questions to test whether the inference has been adequately carried out, and that the associated background knowledge is understood.

3 The Dimensions of Categorization

Inferences are categorized along three dimensions: the *domain* of the background knowledge; the *linguistic significance* i.e. the reason that it is important to make the inference; and the *category of the inference* i.e. the logical structure of the fact being inferred.

3.1 Domain

In the texts we have looked at, the domains fall more or less into six general categories: Spatial and physical knowledge; naive biology; naive psychology; theory of social relations; specialized knowledge such as knowledge about art; and conventions of discourse and narrative (pragmatic knowledge about how texts are structured such as Gricean conventions). Within these we have formulated 21 lower-level categories.

3.2 Linguistic significance

Our analysis of this dimension is, frankly, not yet in a very satisfactory state. There are a few well-defined categories, such as lexical ambiguity, syntactic ambiguity, and coreference resolution; but in many cases our justification for why the inference is important amounts to little more than a description of the kind of fact being inferred plus the claim that it is obviously important to infer facts of that kind. However, we feel that this is an important issue. The set of inferences that one *might* make in reading a text is entirely open-ended; if we want this analysis to be a useful guide to research, we must focus on those that are *important* to make, and it is therefore critical to justify why the inferences are important. We hope that, as our analysis proceeds, the issues here may become clearer.

We have formulated 24 categories found in the texts we have looked at: Abstract frame (in the sense of (Minsky 1975)); characteristic of an entity; clarify misleading syntax; clarify vague expression; coreference resolution; counter argument; ellipsis; explicate causal structure; find case filler; identify entity; interpret non-literal text; lexical disambiguation; motivation analysis; named-entity interpretation; noun-phrase semantic structure; qualification on an event; qualification on a modifier; reference separation; relate example to general description; relation between events; sentiment analysis; semantic disambiguation; source of information; syntactic disambiguation; and temporal sequence.

In some cases, there are linguistic ambiguities that do not end up affecting the meaning of the text. For example, in one sample text about a transit strike, there is a reference to "their busses", where the antecedent of "their" could be either the commuters or the drivers. As it happens, the pronoun can be resolved to the commuters, because of considerations of focus; but it makes no difference how the pronoun is resolved, since it is the same busses either way. We have omitted such cases, since commonsense world knowledge is irrelevant, and the pragmatic discourse knowledge is not very interesting for our purposes.

3.3 Categories of Inference

Here we use a semi-formal structure. The inference is categorized in terms of an operator, which is a relation, and arguments, which are entities. For each operator and argument we specify:

• A general category, from a fairly limited list. For instance `PropertyOf` and `Existence` are categories of relations; `Person` and `Event` are categories of entities.

• A specific value of these categories, with the exception of certain basic relations. For instance in inference 1 above `PropertyOf` has the value `Unlikely` and `Event` has the value `''the impossible''` (in quotes, to emphasize that this is a reference to a phrase in the text). In inference 3, `Existence` has no specific value, and `Event` has the value `Mona Lisa became absent.` These values have no particular structure; they are written in abbreviated English which hopefully is intelligible to the human reader.

• Both relations and entities may have the modifier `Not.` Entities may additionally have the modifier `Multiple.` For example, inference 5 for the above story is the inference that the Mona Lisa was not removed by the museum administration. The operator for this is the relation `Not RoleIn = Actor;` the arguments are `Multiple Person = Administration of Louvre` and `Event = Remove Mona Lisa.`

Currently, the eight categories of entities are `Aspect, Event, Object, Person, Proposition, SpeechAct, State,` and a catch-all `OtherEntity`. The categories of relations are `Authorized, Believe, CausalRelation, ContentOf` (the content of a `SpeechAct`), `Emotion, Ethics, Existence, Goal, Identical, Identify, Motivation, PartOf, Perceive, PropertyOf, QualificationOnProperty, RoleIn, SocialSignificance, SpatialRelation, TemporalRelation,` and `OtherRelation`.

4 Questions and Answers

With each inference, we present one or more multiple-choice question and answer. The corpus indicates which is the right answer.

The question is intended to be based on having read the passage up through the specific text associated with the inference, and are worded accordingly. In some cases, the text further on in the passage gives further information about the answer to the question, and in some cases, later questions implicitly give the answers to earlier ones. Thus, the questions should be viewed as being given sequentially, after the corresponding part of the text has been read.

In formulating the wrong answers, we have been careful to choose them and word them so that they are unequivocally wrong; they do not merely fail to capture the meaning of the text, they are manifestly untrue. In some cases, this limits the precision with which we can test the understanding of the text. Also, some of the wrong answers can be excluded using other kinds of knowledge than we have enumerated under **Background** knowledge needed to corroborate that the correct answer is plausible; it does not capture the knowledge needed to exclude incorrect answers.

It is certainly possible that there are stylistic or other clues in the wording of the question and answer that would allow a program to select the right answer without actually understanding the text. (For example, if an answer contains the word "literally" modifying a word from the text, the answer is probably wrong, since we would only ask about the word in the text if it is being used figuratively.) It would be extremely difficult to prevent this, and we have not made any effort to mitigate it. The set of questions and answer therefore should not be used blindly for evaluation in settings where cheating of this kind is a consideration; challenges of other formats such as the Winograd Schema Challenge (Levesque, Davis, and Morgenstern 2012) are designed to avoid this kind of issue.

5 Difficulties, Obstacles, Limitations

The overriding difficulty of the project is that no aspect is tightly defined and many aspects are very fuzzy indeed. What constitutes an inference, how the annotator individuates two inferences, how the annotator determines the background knowledge, and how the annotator assigns the various categories, are all quite indeterminate. Many of the categories overlap. We have not done studies of interannotator agreement, and our feeling is that, until the theory attains a better-defined state, there would be no point in doing so; the kappa would be too small.

Our hope is that, as we continue the project, analyzing new texts and reviewing our analysis of old texts, the overall structure of the theory will become clearer. One promising sign in that direction is our analysis of the "Categories of inference". In the earlier stages of the project, these were unstructured, in the way that "Linguistic significance" and "Domain" are still unstructured; but, as the project progressed we realized that a better-defined analysis in terms of relations over entities was possible.

Inferences that depend on pragmatic discourse conventions, Gricean rules, and such are in general harder to characterize than those that depends on knowledge of the external world, because these conventions are harder to formulate in anything close to an effectively usable form than facts about the external world.

We illustrate the difficulties that arise with two examples of particular cruxes from the Mona Lisa story above.

"On Tuesday morning . . . she [the Mona Lisa] was gone" is elliptical. It was discovered on Tuesday morning that she

was gone; the theft probably took place Sunday or Monday, and she was actually gone when she was stolen. The inference depends on identifying the point of view, which is an amorphous "public knowledge". The specific knowledge involved is very hard to characterize. (Note that, if it were not for this issue of point of view, Gricean rules would prohibit saying that it was gone on Tuesday if it were already gone on Sunday or Monday.)

"Incredulous reporters" is presumably hyperbolic; the reporters were astonished but they probably did not actually doubt the report that the Mona Lisa had been stolen. Compare, for example, "The Pakistani government claimed that it had no knowledge that Osama bin Laden was living in the compound but many reporters were incredulous," where "incredulous" is meant literally; they believed that the claim was untrue. It is difficult to characterize the background knowledge involved here. What we have settled on is, "The theft of an art work from a museum is not so surprising that a reporter who had recieved a report of this from a reasonable source would seriously doubt its correctness," but that seems awfully specialized.

6 The current state of the corpus; choice of texts; some statistics

Currently, the TACIT corpus includes T texts with S sentences. We have identified Q inferences and B background facts. The corpus is on the open web, at a URL not included here for blind review, and has been placed as supplemental material to this paper for the AAAI reviewers.

The texts that we have used are all news stories. N of these are from the collection of stories examined by LoBue and Yates (2012); one is the first four sentences of the text studied by Gangemi (2013).

Initially we had thought to include as well texts from a biology textbook, and we carried out a preliminary analysis of three such texts. Some interesting categories did emerge in this analysis; for example "find the correspondence between a concrete example and an abstract descrption" as a category in linguistic significance. However, overall the analysis was clearly problematic from the start, and as our analytical framework became better defined, it became impossible. The biology texts were much more challenging that the news texts. In particular, in the analysis of the biology texts, it was extremely difficult, either to break the reasoning down into separate small inferences, or to break the background knowledge down into separate facts. Rather, it seems, large bodies of background knowledge must somehow be applied as a whole to yield rich representations of multiple facts in the text. In fact, in one of our sample texts, on analysis the amount of background knowledge needed for interpretation seemed so large that it was not clear what the "value added" of the text could be; though the text in itself did not by any means seem trivial or platitudinous.

It may also be observed that, among news stories, "human interest" stories, or stories with an unusual twist seem to be more productive than run-of-the-mill stories about bombings, tornados, and so on.

The news stories that we have chosen for analysis are not particularly representative of any category; they are stories that we deliberately chose because they seemed to present interesting issues of interpretation. Also, as we have discussed above, the categories are ill-defined and therefore the assignment of categories is to a significant degree arbitrary. The statistics over domain very largely reflects the choice of texts, which is arbitary; when we included the biology texts, we unsurprisingly had a much larger number of inferences in the "Biology" domain. Therefore the statistics below cannot be taken as in any sense reliable or robust. However, they are to some extent suggestive, so we present them for what they are worth.

TO ADD STATISTICS

7 Related Work

The relation of world knowledge to natural language interpretation has been studied in AI since the seminal works by Charniak (1972) and Schank (1975), and has been revitalized recently by the RTE (Rich Textual Entailment) of (Dagan, GLicksman, and Magnini, 2006),

7.1 LoBue and Yates

The previous work closest to TACIT is a project reported in LoBue and Yates (2012) "Types of Common-Sense Knowledge Neede for Recognizing Textual Entailment". Indeed, in many ways this was the departure point for TACIT, and most of the texts we have analyzed have come out of their collection. A detailed comparison is therefore appropriate.

The ultimate goal of LoBue and Yates was much the same as TACIT; to analyze the kinds of commonsense knowledge used in interpreting text. Their methodology had important points in common with us, and important differences. They examined a corpus of RTE data and manually extracted a set of 108 inferences that required commonsense reasoning to carry out. They then

> created *proofs*, or a step-by-step sketch of the inferences that lead to a decision about entailment of the hypothesis. . . . This labor-intensive process was conducted by one author over more than three months.

These proofs yielded 221 diverse statements of world knowledge. LoBue and Yates divided these into 20 categories of three different flavors:

Form-based categories: Cause-and-effect; preconditions; simultaneous conditions; argument types; prominent relationship; functionality; mutual exclusivity; transitivity.

Content-based category: Arithmetic, geography, public entities, cultural/situational, is member of, has parts, support/opposition, accountability, synechdoche.

Miscellaneous Categories: Probabilistic dependency, omniscience (essentially a closed-world or Gricean condition).

Each background fact was assigned to one category; all but six fit in one category or another.

To test that the categories were well-defined, LoBue and Yates trained some non-experts as annotators. They achieved an overall interannotator agreement of 0.678 in terms of Fleiss' κ, an remarkably impressive figure, considering the task.

Comparing this project to TACIT, we note the following common elements and differences:

Common Elements. The goal is to analyze the background commonsense knowledge used in interpreting text. A collection of background facts for each inference is formulated. LoBue and Yates' "Content-based categories" corresponds in flavor to our "Domains" and their "Form-based categories" corresponds very roughly to our "Categories of Inference", though there is not very much overlap in terms of the specific categories in either case. The number of inferences and of background facts is roughly comparable: LoBue and Yates use 108 inferences and found 221 background facts; we identified ?? inferences and ?? background facts.

Features in LoBue and Yates not in TACIT. LoBue and Yates categorize the background facts rather than the inferences themselves. The background facts have been structured so as to constitute a complete proof in each case; however, though this analysis is certainly heroic, we do not find all of these proofs very convincing (limits of space here precludes a detailed discussion). LoBue and Yates have done tests of interannotator agreement with good outcomes; we have not performed such tests, and, as discussed above, we are very doubtful that they would give good outcomes.

Features in TACIT not in LoBue and Yates. The most important difference between LoBue and Yates and TACIT is in the source of the inferences. In TACIT, as discussed above, we have used a small number of texts, and attempted to identify all the commonsense inferences in those texts. LoBue and Yates, by contrast, used a preexisting data set associated with RTE, which has a single inference associated with each text. On what basis those inferences were selected is not clear; but we do not feel they represent the most interesting, deepest, or more important inferences involved in the text. For example, the RTE inference for the Mona Lisa story quoted above is that Leonardo da Vinci painted the Mona Lisa. A much more striking inference associated with this story is that the Mona Lisa was stolen – a fact, it should be noted, that is never explicitly stated in the above paragraph. Indeed, of the 34 inferences we have analyzed for this story, 12 deal with the theft, the thief, and his actions and motivations. Another example: A text that both we and LoBue and Yates have analyzed is a strange story of how Nadya Suleman, the "Octomom", fired some nurses who were working for free because she thought they were spying on her. The RTE inference associated with this was to infer that Suleman has 14 children in total, from the fact mentioned in the story that she already had six children other than the octuplets. This is hardly the most interesting inference that this story yields.

The result of our more intensive approach to forming the collection of inferences is that many of our inferences are deeper, more complex, and harder to analyze than those in the RTE collection. A much larger fraction of them have to deal with naive psychology, interpersonal relations, and social relations, and a much smaller fraction have to do with relatively straightforward domains such as arithmetic or geography. (LoBue and Yates report that the Geography category accounts for 16.5% of the background facts and Arithmetic for 6.6%; our collection has no inferences in Arithmetic and only one in Geoggraphy.) We do not think that any very strong claims for the representativeness of our collection of inferences can be made; however, we feel fairly confident that they are less unrepresentative than the inferences in the RTE collection.

The other major difference between LoBue and Yates is they have one system of categories, and each fact is assigned to a single category (presumably to facilitate interannotator comparison). By contrast we have three separate dimensions of categories, and inferences may be assigned to multiple categories within each dimension. In particular, they have nothing that corresponds to our "linguistic significance" category, characterizing why the inference is important.

7.2 Other related work

TO BE WRITTEN. RTE generally. SemEval. Gangemi. CYC. Reading evaluation tests.

(Clark 2010), (Gangemi 2012) (Lenat, Prakash, and Shepherd 1985), (Dagan, Glicksman, and Magnini 2006)

8 What has been accomplished?

We have shown that the attempt to identify all the commonsense inferences involved in understanding of simple news stories can yield a diverse, ecologically valid, collection of interesting and deep commonsense inferences. We have made this collection publicly available in a semi-structured form for analysis. This collection has a significantly different flavor from the RTE inferences, or even from the subcollection of RTE inferences that require commonsense reasoning.

We have developed a framework with three dimensions of categorization for characterizing these inferences and the background knowledge they require. We have demonstrated that this framework is reasonably workable for simple news stories. (By contrast, as discussed above, it is not a good framework for the biological texts; so this statement is not vacuous or circular.) The details of the categorization are in a preliminary state; however, as we have progressed in the project, these have become more well-defined and clearer, and we are hopeful that this will continue.

The TACIT project is ongoing and open to the research community; we very much welcome participation from anyone interested.

References

Banko, M., Cafarella, M., Soderland, S., Broadhead, M., & Etzioni, O. (2007). Open Information Extraction from the Web. International Joint Conference on Artificial Intelligence, (pp. 2670-2676).

Brewka, G., Niemelli, I., & Truszczynski, M. (2008). Nonmonotonic Reasoning. In F. Van Harmelen, V. Lifschitz, & B. Porter, Handbook of Knowledge Representation (pp. 239-284). Amsterdam: Elsevier.

Charniak, E. (1972) Toward a Model of Children's Story Comprehension. Ph.D. thesis, MIT.

Clark, P. (2010). RTE5: An informal analysis of some interesting textual entailment examples. Working Note 39.

Cohn, A., & Renz, J. (2007). Qualitative Spatial Reasoning. In F. van Harmelen, V. Lifschitz, & B. Porter, Handbook of Knowledge Representation (pp. 551-596). Elsevier.
Dagan, I., Glicksman, O., & Magnini, B. (2006). The PASCAL Recognising Textual Entailment Challenge. Lecture Notes in Computer Science 3944:177-190.
Gangemi, A. (2013). A comparison of knowledge extraction tools for the semantic web. In The Semantic Web: Semantics and Big Data (pp. 351-366). Springer Berlin Heidelberg.
Havasi, C., Speer, R., Arnold, K., Lieberman, H., Alonso, J., & Moeller, J. (2010). Open Mind Common Sense: Crowdsourcing for common sense. Collaboratively Built Knowledge Sources and AI.
Lenat, D., Prakash, M., & Shepherd, M. (1985). CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. AI Magazine, 6(4), 65-85.
Levesque, H., Davis, E., & Morgenstern, L. (2012). The Winograd Schema Challenge. Principles of Knowledge Representation and Reasoning (KR).
LoBue, P. & Yates, A (2011). Types of Common-Sense Knowledge Needed for Recognizing Textual Entailment. Association for Computational Linguistics (ACL).
Minsky, M. (1975). A Framework for Representing Knowledge. In P. Winston, The Psychology of Computer Vision. New York: McGraw Hill.
Schank, R. (1975). Conceptual Information Processing. Amsterdam: North Holland.