

Types of Common-Sense Knowledge Needed for Recognizing Textual Entailment

Peter LoBue and Alexander Yates

Temple University

Broad St. and Montgomery Ave.

Philadelphia, PA 19130

{peter.lobue,yates}@temple.edu

Abstract

Understanding language requires both linguistic knowledge and knowledge about how the world works, also known as common-sense knowledge. We attempt to characterize the kinds of common-sense knowledge most often involved in recognizing textual entailments. We identify 20 categories of common-sense knowledge that are prevalent in textual entailment, many of which have received scarce attention from researchers building collections of knowledge.

1 Introduction

It is generally accepted that knowledge about how the world works, or common-sense knowledge, is vital for natural language understanding. There is, however, much less agreement or understanding about how to define common-sense knowledge, and what its components are (Feldman, 2002). Existing large-scale knowledge repositories, like Cyc (Guha and Lenat, 1990), OpenMind (Stork, 1999), and Freebase¹, have steadily gathered together impressive collections of common-sense knowledge, but no one yet believes that this job is done. Other databases focus on exhaustively cataloging a specific kind of knowledge — *e.g.*, synonymy and hypernymy in WordNet (Fellbaum, 1998). Likewise, most knowledge extraction systems focus on extracting one specific kind of knowledge from text, often factual relationships (Banko et al., 2007; Suchanek et al., 2007; Wu and Weld, 2007), although other specialized extraction techniques exist as well.

¹<http://www.freebase.com/>

If we continue to build knowledge collections focused on specific types, will we collect a sufficient store of common sense knowledge for understanding language? What kinds of knowledge might lie outside the collections that the community has focused on building? We have undertaken an empirical study of a natural language understanding task in order to help answer these questions. We focus on the Recognizing Textual Entailment (RTE) task (Dagan et al., 2006), which is the task of recognizing whether the meaning of one text, called the Hypothesis (H), can be inferred from another, called the Text (T). With the help of five annotators, we have investigated the RTE-5 corpus to determine the types of knowledge involved in human judgments of RTE. We found 20 distinct categories of common-sense knowledge that featured prominently in RTE, besides linguistic knowledge, hyponymy, and synonymy. Inter-annotator agreement statistics indicate that these categories are well-defined. Many of the categories fall outside of the realm of all but the most general knowledge bases, like Cyc, and differ from the standard relational knowledge that most automated knowledge extraction techniques try to find.

The next section outlines the methodology of our empirical investigation. Section 3 presents the categories of world knowledge that we found were most prominent in the data. Section 4 discusses empirical results of our survey.

2 Methodology

We follow the methodology outlined in Sammons *et al.* (2010), but unlike theirs and other previous studies (Clark et al., 2007), we concentrate on the world

#56 - ENTAILMENT

T: (CNN) Nadya Suleman, the Southern California woman who gave birth to octuplets in January, [...] She now has four of the octuplets at home, along with her six other children.

1) “octuplets” are 8 children (*definitional*)

2) $8 + 6 = 14$ children (*arithmetic*)

H: Nadya Suleman has 14 children.

Figure 1: An example RTE label, Text, a condensed “proof” (with knowledge categories for the background knowledge) and Hypothesis.

knowledge rather than linguistic knowledge required for RTE. First, we manually selected a set of RTE data that could not be solved using linguistic knowledge and WordNet alone. We then sketched step-by-step inferences needed to show ENTAILMENT or CONTRADICTION of the hypothesis. We identified prominent categories of world knowledge involved in these inferences, and asked five annotators to label the knowledge with the different categories. We judge the well-definedness of the categories by inter-annotator agreement, and their relative importance according to frequency in the data.

To select an appropriate subset of the RTE data, we discarded RTE pairs labeled as UNKNOWN. We also discarded RTE pairs with ENTAILMENT and CONTRADICTION labels, if the decision relies mostly or entirely on a combination of linguistic knowledge, coreference decisions, synonymy, and hypernymy. These phenomena are well-known to be important to language understanding and RTE (Mirkin et al., 2009; Roth and Sammons, 2007). Many synonymy and hypernymy databases already exist, and although coreference decisions may themselves depend on world knowledge, it is difficult to separate the contribution of world knowledge from the contribution of linguistic cues for coreference. Some sample phenomena that we explicitly chose to disregard include: knowledge of syntactic variations, verb tenses, apposition, and abbreviations. From the 600 T and H pairs in RTE-5, we selected 108 that did not depend only on these phenomena.

For each of the 108 pairs in our data, we created *proofs*, or a step-by-step sketch of the inferences that lead to a decision about entailment of the hypothesis.

Figure 1 shows a sample RTE pair and (condensed) proof. Each line in the proof indicates either a new piece of background knowledge brought to bear, or a *modus ponens* inference from the information in the text or previous lines of the proof. This labor-intensive process was conducted by one author over more than three months. Note that the proofs may not be the only way of reasoning from the text to an entailment decision about the hypothesis, and that alternative proofs might require different kinds of common-sense knowledge. This caveat should be kept in mind when interpreting the results, but we believe that by aggregating over many proofs, we can counter this effect.

We created 20 categories to classify the 221 diverse statements of world knowledge in our proofs. These categories are described in the next section.² In some cases, categories overlap (*e.g.*, “Canberra is part of Australia” could be in the *Geography* category or the *part of* category). In cases where we foresaw the overlaps, we manually specified which category should take precedence; in the above example, we gave precedence to the *Geography* category, so that statements of this kind would all be included under *Geography*. This approach has the drawback of biasing somewhat the frequencies in our data set towards the categories that take precedence. However, this simplification significantly reduces the annotation effort of our survey participants, who already face a complicated set of decisions.

We evaluate our categorization to determine how well-defined and understandable the categories are. We conducted a survey of five undergraduate students, who were all native English speakers but otherwise unfamiliar with NLP. The 20 categories were explained using fabricated examples (not part of the survey data). Annotators kept these fabricated examples as references during the survey. Each annotator labeled each of the pieces of world knowledge from the proofs using one of the 20 categories. From this data we calculate Fleiss’s κ for inter-annotator agreement³ in order to measure how well-defined the categories are. We compute κ once over all ques-

²The RTE pairs, proofs, and category judgments from our study are available at <http://www.cis.temple.edu/~yates/data/rte-study-data.zip>

³Fleiss’s κ handles more than two annotators, unlike the more familiar Cohen’s κ .

tions and all categories. Separately, we also compute κ once for each category C , by treating all annotations for categories $C' \neq C$ as the same.

3 Categories of Knowledge

By manual inspection, we arrived at the following 20 prominent categories of world knowledge in our subset of the RTE-5 data. For each category, we give a brief definition and example, along with the ID of an RTE pair whose proof includes the example. Our categories can be loosely organized into form-based categories and content-based categories. Note that, as with most common-sense knowledge, our examples are intended as rules that are usually or typically true, rather than categorically or universally true.

3.1 Form-based Categories

The following categories are defined by how the knowledge can be described in a representation language, such as logic.

1. Cause and Effect: Statements in this category require that a predicate p holds true after an event or action A .

#542: Once a person is welcomed into an organization, they belong to that organization.

2. Preconditions: For a given action or event A at time t , a precondition p is a predicate that must hold true of the world before time t , in order for A to have taken place.

#372: To become a naturalized citizen of a place, one must not have been born there.

3. Simultaneous Conditions: Knowledge in this category indicates that a predicate p must hold true at the same time as an event or second predicate p' .

#240: When a person is an employee of an organization, that organization pays his or her salary.

4. Argument Types: Knowledge in this category specifies the *types* or selectional preferences for arguments to a relationship.

#311: The type of thing that adopts children is the type *person*.

5. Prominent Relationship: Texts often specify that there exists some relationship between two entities, without specifying which relationship. Knowledge in this category specifies which relationship is most likely, given the types of the entities involved.

#42: If a painter is related to a painting somehow

(e.g., “da Vinci’s *Mona Lisa*”), the painter most likely *Painted* the painting.

6. Definition: Any explanation of a word or phrase.

#163: A “seat” is an object which holds one person.

7. Functionality: This category lists relationships R which are *functional*; i.e., $\forall x,y,y' R(x,y) \wedge R(x,y') \Rightarrow y = y'$.

#493: *fatherOf* is functional — a person can have only one father.

8. Mutual Exclusivity: Related to functionality, mutual exclusivity knowledge indicates types of things that do not participate in the same relationship.

#229: Government and media sectors usually do not employ the same person at the same time.

9. Transitivity: If we know that R is transitive, and that $R(a,b)$ and $R(b,c)$ are true, we can infer that $R(a,c)$ is true.

#499: The *supports* relation is transitive. Thus, because Putin supports the United Russia party, and the United Russia party supports Medvedev, we can infer that Putin supports Medvedev.

3.2 Content-based Categories

The following categories are defined by the content, topic, or domain of the knowledge in them.

10. Arithmetic: This includes addition and subtraction, as well as comparisons and rounding.

#609: 115 passengers + 6 crew = 121 people

11. Geography: This includes knowledge such as “Australia is a place,” “Sydney is in Australia,” and “Canberra is the capital of Australia.”

12. Public Entities: This category is for well-known properties of highly-recognizable named-entities.

#142: Berlusconi is prime minister of Italy.

13. Cultural/Situational: This category includes knowledge of or shared by a particular culture.

#207: A “half-hour drive” is “near.”

14. is member of: Statements of this category indicate that an entity belongs to a larger organization.

#374: A minister is part of the government.

15. has parts: This category expresses what components an object or situation is comprised of.

#463: Forests have trees.

16. Support/Opposition: This includes knowledge of the kinds of actions or relationships toward X that indicate positive or negative feeling toward X .

#357: P finds $X \Rightarrow P$ supports X

17. Accountability: This includes any knowledge that is helpful for determining who or what is responsible for an action or event.

#158: A nation’s military is responsible for that nation’s bombings.

18. Synecdoche: Synecdoche is knowledge that a person or thing can represent or speak for an organization or structure he or she is a part of.

#410: The president of Russia represents Russia.

3.3 Miscellaneous Categories

19. Probabilistic Dependency: Multiple phrases in the text may contribute to the hypothesis being more or less likely to be true, although each phrase on its own might not be sufficient to support the hypothesis. Knowledge in this category indicates that these separate pieces of evidence can combine in a probabilistic, noisy-or fashion to increase confidence in a particular inference.

#437: Stocks on the “Nikkei 225” exchange and Toyota’s stock both fell, which independently suggest that Japan’s economy might be struggling, but in combination they are stronger evidence that Japan’s economy is floundering.

20. Omniscience: Certain RTE judgments are only possible if we assume that the text includes all information pertinent to the story, so that we may discredit statements that were not mentioned.

#208: T states that “Fitzpatrick pleaded guilty to fraud and making a false report.” H, which is marked as a CONTRADICTION, states that “Fitzpatrick is accused of robbery.” In order to prove the falsehood of H, we had to assume that no charges were made other than the ones described in T.

4 Results and Discussion

Our headline result is that the above twenty categories overall are well-defined, with a Fleiss’s κ score of 0.678, and that they cover the vast majority of the world knowledge used in our proofs. This has important implications, as it suggests that concentrating on collecting these kinds of world knowledge will make a large difference to RTE, and hopefully to language understanding in general. Naturally, more studies of this issue are warranted for validation.

Many of the categories — has parts, member of, geography, cause and effect, public entities, and

Category	Occurrences	κ
Functionality	19.2 (8.7%)	0.663
Definitions	17.2 (7.8%)	0.633
Preconditions	15.8 (7.1%)	0.775
Cause and Effect	10.8 (4.9%)	0.591
Prominent Relationship	8.4 (3.8%)	0.145
Argument Types	6.8 (3.1%)	0.180
Simultaneous Conditions	6.2 (2.8%)	0.203
Mutual Exclusivity	6 (2.7%)	0.640
Transitivity	3 (1.4%)	0.459
Geography	36.4 (16.5%)	0.927
Support/Opposition	14.6 (6.6%)	0.684
Arithmetic	13.4 (6.1%)	0.968
is member of	11.6 (5.2%)	0.663
Synecdoche	9.8 (4.4%)	0.829
has parts	8.8 (4.0%)	0.882
Accountability	7.2 (3.3%)	0.799
Cultural/Situational	4.6 (2.1%)	0.267
Public Entities	3.2 (1.4%)	0.429
Omniscience	7.2 (3.3%)	0.828
Probabilistic Dependency	4.8 (2.2%)	0.297
All	215 (97%)	0.678

Table 1: **Frequency and inter-annotator agreement for each category of world knowledge in the survey.** Frequencies are averaged over the five annotators, and agreement is calculated using Fleiss’s κ .

support/opposition — will be familiar to NLP researchers from resources like WordNet, gazetteers, and text mining projects for extracting causal knowledge, properties of named entities, and opinions. Yet these familiar categories make up only about 40% of the world knowledge used in our proofs. Common knowledge types, like definitional knowledge, arithmetic, and accountability, have for the most part been ignored by research on automated knowledge collection. Others have only earned very scarce and recent attention, like preconditions (Sil et al., 2010) and functionality (Ritter et al., 2008).

Several interesting form-based categories, including **Prominent relationships**, **Argument Types**, and **Simultaneous Conditions**, had quite low inter-annotator agreement. We continue to believe that these are well-defined categories, and suspect that

further studies with better training of the annotators will support this. One issue during annotation was that certain pieces of knowledge could be labeled as a content category or a form category, and instructions may not have been clear enough on which is appropriate under these circumstances. Nevertheless, considering the number of annotators and the uneven distribution of data points across the categories (both of which tend to decrease κ), κ scores are overall quite high.

In an effort to discover if some of the categories overlap enough to justify combining them into a single category, we tried combining categories which annotators frequently confused with one another. While we could not find any combination that significantly improved the overall κ score, several combinations provided minor improvements. As an example of a merge that failed, we tried merging **Argument Types** and **Mutual Exclusivity**, with the idea that if a system knows about the selectional preferences of different relationships, it should be able to deduce which relationships or types are mutually exclusive. However, the κ score for this combined category was 0.410, significantly below the κ of 0.640 for **Mutual Exclusivity** on its own. One merge that improves κ is a combination of **Prominent Relationship** with **Argument Types** (combined κ of 0.250, as compared with 0.145 for **Prominent Relationship** and 0.180 for **Argument Types**). However, we believe this is due to unclear wording in the proofs, rather than a real overlap between the two categories. For instance, “Painters paint paintings” is an example of the **Prominent Relationship** category, and it looks very similar to the **Argument Types** example, “People adopt children.” The knowledge in the first case is more properly described as, “If there exists an unspecified relationship R between a painter and a painting, then R is the relationship ‘painted’.” In the second case, the knowledge is more properly described as, “If x participates in the relationship ‘adopts children’, then x is of type ‘person’.” Stated in this way, these kinds of knowledge look quite different. If one reads our proofs from start to finish, the flow of the argument indicates which of these forms is intended, but for annotators quickly reading through the proofs, the two kinds of knowledge can look superficially very similar, and the annotators can become confused.

The best category combination that we discovered is a combination of **Functionality** and **Mutual Exclusivity** (combined κ of 0.784, compared with 0.663 for **Functionality** and 0.640 for **Mutual Exclusivity**). This is a potentially valid alternative to our classification of the knowledge. Functional relationships R imply that if x and x' have different values y and y' , then x and x' must be distinct, or mutually exclusive. We intended that **Mutual Exclusivity** apply to sets rather than individual items, but annotators apparently had trouble distinguishing between the two categories, so in future we may wish to revise our set of categories. Further surveys would be required to validate this idea.

The 20 categories of knowledge covered 215 (97%) of the 221 statements of world knowledge in our proofs. Of the remaining 6 statements, two were from recognizable categories, like knowledge for temporal reasoning (**#355**) and an application of the frame axiom (**#265**). We left these out of the survey to cut down on the number of categories that annotators had to learn. The remaining four statements were difficult to categorize at all. For instance, **#177**: “Motorcycle manufacturers often sponsor teams in motorcycle sports.” The other three of these difficult-to-categorize statements came from proofs for **#265**, **#336**, and **#432**. We suspect that if future studies analyze more data for common-sense knowledge types, more categories will emerge as important, and more facts that lie outside of recognizable categories will also appear. Fortunately, however, it appears that at least a very large fraction of common-sense knowledge can be captured by the sets of categories we describe here. Thus these categories serve to point out promising areas for further research in collecting common-sense knowledge.

References

- M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. 2007. Open information extraction from the web. In *IJCAI*.
- Peter Clark, William R. Murray, John Thompson, Phil Harrison, Jerry Hobbs, and Christiane Fellbaum. 2007. On the role of lexical and world knowledge in rte3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, RTE '07, pages 54–59, Morristown, NJ, USA. Association for Computational Linguistics.

- I. Dagan, O. Glickman, and B. Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. *Lecture Notes in Computer Science*, 3944:177–190.
- Richard Feldman. 2002. *Epistemology*. Prentice Hall.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.
- R.V. Guha and D.B. Lenat. 1990. Cyc: a mid-term report. *AI Magazine*, 11(3).
- V. Vydiswaran M. Sammons and D. Roth. 2010. Ask not what textual entailment can do for you... In *Proc. of the Annual Meeting of the Association of Computational Linguistics (ACL)*, Uppsala, Sweden, 7. Association for Computational Linguistics.
- Shachar Mirkin, Ido Dagan, and Eyal Shnarch. 2009. Evaluating the inferential utility of lexical-semantic resources. In *EACL*.
- Alan Ritter, Doug Downey, Stephen Soderland, and Oren Etzioni. 2008. It's a contradiction — No, it's not: A case study using functional relations. In *Empirical Methods in Natural Language Processing*.
- Dan Roth and Mark Sammons. 2007. Semantic and logical inference model for textual entailment. In *Proceedings of ACL-WTEP Workshop*.
- Avirup Sil, Fei Huang, and Alexander Yates. 2010. Extracting action and event semantics from web text. In *AAAI Fall Symposium on Common-Sense Knowledge (CSK)*.
- D. G. Stork. 1999. The OpenMind Initiative. *IEEE Expert Systems and Their Applications*, 14(3):19–20.
- Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on the World Wide Web (WWW)*.
- Fei Wu and Daniel S. Weld. 2007. Automatically semantifying wikipedia. In *Sixteenth Conference on Information and Knowledge Management (CIKM-07)*.