Discovering Commonsense Entailment Rules Implicit in Sentences

Jonathan Gordon
Department of Computer Science
University of Rochester
Rochester, NY, USA
jgordon@cs.rochester.edu

Lenhart K. Schubert
Department of Computer Science
University of Rochester
Rochester, NY, USA
schubert@cs.rochester.edu

Abstract

Reasoning about ordinary human situations and activities requires the availability of diverse types of knowledge, including expectations about the probable results of actions and the lexical entailments for many predicates. We describe initial work to acquire such a collection of conditional (if–then) knowledge by exploiting presuppositional discourse patterns (such as ones involving 'but', 'yet', and 'hoping to') and abstracting the matched material into general rules.

1 Introduction

We are interested, ultimately, in enabling an inference system to reason forward from facts as well as backward from goals, using lexical knowledge together with world knowledge. Creating appropriate collections of general world knowledge to support reasoning has long been a goal of researchers in Artificial Intelligence. Efforts in information extraction, *e.g.*, Banko *et al*. (2007), have focused on learning base facts about specific entities (such as that Barack Obama is president), and work in knowledge extraction, *e.g.*, Van Durme and Schubert (2008), has found generalizations (such as that a president may make a speech). While the latter provides a basis for possibilistic forward inference (Barack Obama probably makes a speech at least occasionally) when its meaning is sharpened (Gordon and Schubert, 2010), these resources don't provide a basis for saying what we might expect to happen if, for instance, someone crashes their car.

That the driver in a car crash might be injured and the car damaged is a matter of common sense, and, as such, is rarely stated directly. However, it can be found in sentences where this expectation is disconfirmed: 'Sally crashed her car into a tree, but she wasn't hurt.' We have been exploring the use of lexico-syntactic discourse patterns indicating disconfirmed expectations, as well as people's goals ('Joe apologized repeatedly, hoping to be forgiven'). The resulting rules, expressed at this point in natural language, are a first step toward obtaining classes of general conditional knowledge typically not obtained by other methods.

2 Related Work

One well-known approach to conditional knowledge acquisition is that of Lin and Pantel (2001), where inference rules are learned using distributional similarity between dependency tree paths. These results include entailment rules like 'x is the author of $y \Leftrightarrow x$ wrote y' (which is true provided x is a literary work) and less dependable ones like 'x caused $y \Leftrightarrow y$ is blamed on x'. This work was refined by Pantel *et al*. (2007) by assigning the x and y terms semantic types (*inferential selectional preferences* – ISP) based on lexical abstraction from empirically observed argument types. A limitation of the approach is that the conditional rules obtained are largely limited to ones expressing some rough synonymy or similarity relation. Pekar (2006) developed related methods for learning the implications of an event based on the regular co-occurrence of two verbs within "locally coherent text", acquiring rules like 'x was appointed as y' suggests that 'x became y', but, as in DIRT, we lack information about the types of x and y, and only acquire binary relations.

Girju (2003) applied Hearst's (1998) procedure for finding lexico-syntactic patterns to discover causal relations between nouns, as in 'Earthquakes generate tsunami'. Chklovski and Pantel (2004) used pat-

```
(S < (NP $. (VP < (/,/ $. (S < (VP< (VBG<hoping)< (S< (VP< TO)))))))))
(S < (NP $. (VP < ((CC < but) $.. (VP < (AUX < did) < (RB </n['o]t/))))))
(S < (NP $. (VP < (AUX $. (ADJP < (JJ $. ((CC < /(but|yet)/) $. JJ)))))))
(S < (NP $. (VP < (/,/ $. (S < (VP < ((VBG < expecting) $.
                                      (S < (VP < TO)))))))))
```

Figure 1: Examples of TGrep2 patterns for finding parse tree fragments that might be abstracted to inference rules. See Rohde (2001) for an explanation of the syntax.

terns like '*x*-ed by *y*-ing' ('obtained by borrowing') to get co-occurrence data on candidate pairs from the Web. They used these co-occurrence counts to obtain a measure of mutual information between pairs of verbs, and hence to assess the strengths of the relations. A shortcoming of rules obtained in this way is their lack of detailed predicative structure. For inference purposes, it would be insufficient to know that 'crashes cause injuries' without having any idea of what is crashing and who or what is being injured.

Schoenmackers *et al*. (2010) derived first-order Horn clauses from the tuple relations found by TEXTRUNNER (Banko *et al*., 2007). Their system produces rules like 'IsHeadquarteredIn(Company, State) :- IsBasedIn(Company, City) ∧ IsLocatedIn(City, State)', which are intended to improve inference for question-answering. A limitation of this approach is that, operating on the facts discovered by an information extraction system, it largely obtains relations among simple attributes like locations or roles rather than consequences or reasons.

3 Method

Our method first uses TGrep2 (Rohde, 2001) to find parse trees matching hand-authored lexico-syntactic patterns, centered around certain pragmatically significant cue words such as 'hoping to' or 'but didn't'. Some of the search patterns are in Figure 1. While we currently use eight query patterns, future work may add rules to cover more constructions.

The matched parse trees are filtered to remove those unlikely to produce reasonable results, such as those containing parentheses or quoted utterances, and the trees are preprocessed in a top-down traversal to rewrite or remove constituents that are usually extraneous. For instance, the parse tree for

> *The next day he and another Bengali boy who lives near by* [sic] *chose another way home, hoping to escape the attackers.*

is preprocessed to

> People chose another way home, hoping to escape the attackers.

Examples of the preprocessing rules include removing interjections (INTJ) and some prepositional phrases, heuristically turning long expressions into keywords like 'a proposition', abstracting named entities, and reordering some sentences to be easier to process. *E.g.*, 'Fourteen inches from the floor it's supposed to be' is turned to 'It's supposed to be fourteen inches from the floor'.

The trees are then rewritten as conditional expressions based on which semantic pattern they match, as outlined in the following subsections. The sample sentences are from the Brown Corpus (Kučera and Francis, 1967) and the British National Corpus (BNC Consortium, 2001), and the rules are those derived by our current system.

3.1 Disconfirmed Expectations

These are sentences where 'but' or 'yet' is used to indicate that the expected inference people would make does not hold. In such cases, we want to flip the polarity of the conclusion (adding or removing 'not' from the output) so that the expectation is confirmed. For instance, from

> *The ship weighed anchor and ran out her big guns, but did not fire a shot.*

we get that the normal case is the opposite:

> If a ship weighs anchor and runs out her big guns, then it may fire a shot.

Or for two adjectives, 'She was poor but proud':

> If a female is poor, then she may not be proud.

3.2 Contrasting Good and Bad

A different use of 'but' and 'yet' is to contrast something considered good with something considered bad, as in 'He is very clever but eccentric':

> If a male is very clever,
> then he may be eccentric.

If we were to treat this as a case of disconfirmed expectation as above, we would have claimed that 'If a male is very clever, then he may not be eccentric'. To identify this special use of 'but', we consult a lexicon of sentiment annotations, SentiWordNet (Baccianella *et al*., 2010). Finding that 'clever' is positive while 'eccentric' is negative, we retain the surface polarity in this case.

For sentences with full sentential complements for 'but', recognizing good and bad items is quite difficult, more often depending on pragmatic information. For instance, in

> *Central government knew this would happen but did not want to admit to it in its plans.*

knowing something is generally good while being unwilling to admit something is bad. At present, we don't deal with these cases.

3.3 Expected Outcomes

Other sentences give us a participant's intent, and we just want to abstract sufficiently to form a general rule:

> *He stood before her in the doorway, evidently expecting to be invited in.*
>
> If a male stands before a female in the doorway, then he may expect to be invited in.

When we abstract from named entities (using a variety of hand-built gazetteers), we aim low in the hierarchy:

> *Elisabeth smiled, hoping to lighten the conversational tone and distract the Colonel from his purpose.*
>
> If a female smiles, then she may hope to lighten the conversational tone.

While most general rules about 'a male' or 'a female' could instead be about 'a person', there are ones that can't, such as those about giving birth. We leave the raising of terms for later work, following Van Durme *et al*. (2009).

4 Evaluation

Development was based on examples from the (hand-parsed) Brown Corpus and the (machine-parsed) British National Corpus, as alluded to above. These corpora were chosen for their broad coverage of everyday situations and edited writing.

As the examples in the preceding subsections indicate, rules extracted by our method often describe complex consequences or reasons, and subtle relations among adjectival attributes, that appear to be quite different from the kinds of rules targeted in previous work (as discussed earlier, or at venues such as that of (Sekine, 2008)). While we would like to evaluate the discovered rules by looking at inferences made with them, that must wait until logical forms are automatically created; here we judge the rules themselves.

> The statement above is a reasonably clear, entirely plausible, generic claim and seems neither too specific nor too general or vague to be useful:
>
> 1. I agree.
> 2. I lean towards agreement.
> 3. I'm not sure.
> 4. I lean towards disagreement.
> 5. I disagree.

Figure 2: Instructions for judging of unsharpened factoids.

Judge 1	*Judge 2*	*Correlation*
1.84	2.45	0.55

Table 1: Average ratings and Pearson correlation for rules from the personal stories corpus. Lower ratings are better; see Fig. 2.

For evaluation, we used a corpus of personal stories from weblogs (Gordon and Swanson, 2009), parsed with a statistical parser (Charniak, 2000). We sampled 100 output rules and rated them on a scale of 1–5 (1 being best) based on the criteria in Fig. 2. To decide if a rule meets the criteria, it is helpful to imagine a dialogue with a computer agent. Told an instantiated form of the antecedent, the agent asks for confirmation of a potential conclusion. *E.g.*, for

> If attacks are brief,
> then they may not be intense,

the dialogue would go:

> "The attacks (on Baghdad) were brief."
> "So I suppose they weren't intense, were they?"

If this is a reasonable follow-up, then the rule is probably good, although we also disprefer very unlikely antecedents – rules that are vacuously true.

As the results in Table 1 and Fig. 3 indicate, the overall quality of the rules learned is good but there is room for improvement. We also see a rather low correlation between the ratings of the two judges, indicating the difficulty of evaluating the quality of the rules, especially since their expression in natural language (NL) makes it tempting to "fill in the blanks" of what we understand them to mean. We hypothesize that the agreement between judges will be higher for rules in logical form, where malformed output is more readily identified – for instance, there is no guessing about coreference or attachment.

Rules that both judges rated favorably (1) include:

> If a pain is great, it may not be manageable.
>
> If a person texts a male, then he-or-she may get a reply.

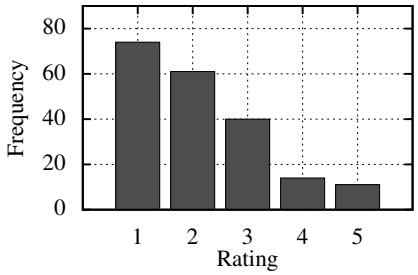

Figure 3: Counts for how many rules were assigned each rating by judges. Lower ratings are better; see Fig. 2.

> If a male looks around, then he may hope to see someone.
>
> If a person doesn't like some particular store, then he-or-she may not keep going to it.

While some bad rules come from parsing or processing mistakes, these are less of a problem than the heavy tail of difficult constructions. For instance, there are idioms that we want to filter out (*e.g.*, 'I'm embarrassed but...') and other bad outputs show context-dependent rather than general relations:

> If a girl sits down in a common room, then she may hope to avoid some pointless conversations.

The sitting-down may not have been *because* she wanted to avoid conversation but because of something prior.

It's difficult to compare our results to other systems because of the differences of representation, types of rules, and evaluation methods. ISP's best performing method (ISP.JIM) achieves 0.88 specificity (defined as a filter's probability of rejecting incorrect inferences) and 0.53 accuracy. While describing their SHERLOCK system, Schoenmackers *et al*. (2010) argue that "the notion of 'rule quality' is vague except in the context of an application" and thus they evaluate the Horn clauses they learn in the context of the HOLMES inference-based QA system, finding that at precision 0.8 their rules allow the system to find twice as many correct facts. Indeed, our weak rater agreement shows the difficulty of judging rules on their own, and future work aims to evaluate rules extrinsically.

5 Conclusion and Future Work

Enabling an inference system to reason about common situations and activities requires more types of general world knowledge and lexical knowledge than are currently available or have been targeted by previous work. We've suggested an initial approach to acquiring rules describing complex consequences or reasons and subtle relations among adjectival attributes: We find possible rules by looking at interesting discourse patterns and rewriting them as conditional expressions based on semantic patterns.

A natural question is why we don't use the machine-learning/bootstrapping techniques that are common in other work on acquiring rules. These techniques are particularly successful when (a) they are aimed at finding fixed types of relationships, such as hyponymy, near-synonymy, part-of, or causal relations between pairs of lexical items (often nominals or verbs); and (b) the fixed type of relationship between the lexical items is hinted at sufficiently often either by their co-occurrence in certain local lexico-syntactic patterns, or by their occurrences in similar sentential environments (distributional similarity). But in our case, (a) we are looking for a broad range of (more or less strong) consequence relationships, and (b) the relationships are between entire clauses, not lexical items. We are simply not likely to find multiple occurrences of the same pair of clauses in a variety of syntactic configurations, all indicating a consequence relation – you're unlikely to find multiple redundant patterns relating clauses, as in 'Went up to the door but didn't knock on it'.

There is more work to be done to arrive at a reliable, inference-ready knowledge base of such rules. The primary desideratum is to produce a logical representation for the rules such that they can be used in the EPILOG reasoner (Schubert and Hwang, 2000). Computing logical forms (as, *e.g.*, in Bos (2008)) and then deriving logically formulated rules from these rather than deriving sentential forms directly from text should also allow us to be more precise about dropping modifiers, reshaping into generic present tense from other tenses, and other issues that affect the quality of the statements. We have a preliminary version of a logical form generator that derives LFs from TreeBank parses that can support this direction. Further filtering techniques (based both on the surface form and the logical form) should keep the desired inference rules while improving quality.

Acknowledgements

This work was supported by NSF grants IIS-1016735 and IIS-0916599, and ONR STTR subcontract N00014-10-M-0297.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proc. of the Seventh Conference on International Language Resources and Evaluation (LREC'10)*.

Michele Banko, Michael J. Cafarella, Stephen Soderland, Matt Broadhead, and Oren Etzioni. 2007. Open information extraction from the Web. In *Proc. of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)*.

BNC Consortium. 2001. The British National Corpus, v.2. Distributed by Oxford University Computing Services.

Johan Bos. 2008. Wide-coverage semantic analysis with Boxer. In *Proc. of the Symposium on Semantics in Text Processing (STEP 2008)*.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proc. of the First Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL 2000)*, pages 132–139.

Timothy Chklovski and Patrick Pantel. 2004. VerbOcean: Mining the Web for fine-grained semantic verb relations. In *Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proc. of the ACL 2003 Workshop on Multilingual Summarization and Question Answering – Machine Learning and Beyond*.

Jonathan Gordon and Lenhart K. Schubert. 2010. Quantificational sharpening of commonsense knowledge. In *Proc. of the AAAI 2010 Fall Symposium on Commonsense Knowledge*.

Andrew Gordon and Reid Swanson. 2009. Identifying personal stories in millions of weblog entries. In *Proc. of the Third International Conference on Weblogs and Social Media (ICWSM), Data Challenge Workshop*.

Marti Hearst. 1998. Automated discovery of WordNet relations. In Christiane Fellbaum, editor, *WordNet: An Electronic Lexical Database and Some of Its Applications*. MIT Press.

Henry Kučera and W. N. Francis. 1967. *Computational Analysis of Present-Day American English*. Brown University Press.

Dekang Lin and Patrick Pantel. 2001. DIRT: Discovery of inference rules from text. In *Proc. of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

Patrick Pantel, Rahul Bhagat, Timothy Chklovski, and Eduard Hovy. 2007. ISP: Learning inferential selectional preferences. In *Proc. of NAACL-HLT 2007*.

Viktor Pekar. 2006. Acquisition of verb entailment from text. In *Proc. of HLT-NAACL 2006*.

Doug Rohde. 2001. TGrep2 manual. Unpublished manuscript, Brain & Cognitive Science Department, MIT.

Stefan Schoenmackers, Oren Etzioni, Daniel S. Weld, and Jesse Davis. 2010. Learning first-order horn clauses from Web text. In *Proc. of EMNLP 2010*.

Lenhart K. Schubert and Chung Hee Hwang. 2000. Episodic Logic Meets Little Red Riding Hood: A comprehensive, natural representation for language understanding. In L. Iwanska and S. C. Shapiro, editors, *Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language*. MIT/AAAI Press.

Satoshi Sekine, editor. 2008. *Notebook of the NSF Symposium on Semantic Knowledge Discovery, Organization, and Use*. New York University, 14–15 November.

Benjamin Van Durme and Lenhart K. Schubert. 2008. Open knowledge extraction through compositional language processing. In *Proc. of the Symposium on Semantics in Text Processing (STEP 2008)*.

Benjamin Van Durme, Phillip Michalak, and Lenhart K. Schubert. 2009. Deriving Generalized Knowledge from Corpora using WordNet Abstraction. In *Proc. of EACL 2009*.