Reporting Bias and Knowledge Acquisition

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

Benjamin Van Durme
HLTCOE
Johns Hopkins University
Baltimore, MD, USA
vandurme@cs.jhu.edu

ABSTRACT

Much work in knowledge extraction from text tacitly assumes that the frequency with which people write about actions, outcomes, or properties is a reflection of real-world frequencies or the degree to which a property is characteristic of a class of individuals. In this paper, we question this idea, examining the phenomenon of *reporting bias* and the challenge it poses for knowledge extraction. We conclude with discussion of approaches to learning commonsense knowledge from text in spite of this distortion.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning—*Knowledge Acquisition*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text Analysis*

General Terms

Theory, Measurement

1. INTRODUCTION

In Artificial Intelligence, it seems that the human-like understanding and reasoning required for problems such as question-answering, recognizing textual entailment, and planning depends on access to large amounts of general world knowledge. The difficulty of accumulating such a collection is known as the *knowledge acquisition bottleneck*. While there have been attempts to manually engineer this knowledge (most notably the Cyc project [21]) or to solicit it directly from online crowds (e.g., the Open Mind Initiative [28]), the dominant approach is to mine knowledge from the vast amounts of text available in electronic form.

A knowledge-extraction system can look for explicit assertions of general knowledge or knowledge implicit in recurrent patterns of predication and modification; or it can abstract general claims from collections of specific instances. Regardless of the modus operandi, it is necessary to distinguish knowledge about what *normally* holds in the world from what is possible but atypical or output that is simply inaccurate. The latter doesn't necessarily indicate a failure of the system to learn from its input; a text may be inaccurate or fantastical. For instance, a system that reads Web text and believes with equal confidence everything it reads would learn both *The Earth revolves around the Sun* and *The Sun revolves around the Earth*. Indeed, knowledge-acquisition systems such as Knext will find both of these.[1]

To distinguish these claims, it is typical to take an inductive view, with textual references serving as evidence. This is intuitively reasonable: The more often we read something, the more likely it is to reflect the truth of the real world. Over a large collection of texts, Knext's heliocentric claim is learned from 107 textual references, while its geocentric claim is only learned from 50.

However, on closer examination, the frequency with which situations of a certain type are described in text do not necessarily correspond to their relative likelihood in the world, or even the subjective frequencies captured in human beliefs. For instance, from the same texts, Knext learns over a million times that *A person may have eyes*, but fewer than 1,500 times that *A person may have a spleen*. While eyes are discussed frequently, many other body parts are not – but this doesn't mean they're any less common in people. We will refer to this potential discrepancy between reality and its description in text as *reporting bias* [30].

For knowledge-extraction, we are interested in reporting bias as it relates to the frequency of events or actions occurring, the frequency of specific outcomes occurring, and the frequency of properties. If our textual examples are not representative of reality, then claims induced from them are likely to be inaccurate. For instance, according to Doug Lenat, at one point Cyc "concluded that everyone born before 1900 was famous, because all the people that it knew about and who lived in earlier times were famous people." [22]

In Section 2, we present evidence of reporting bias by contrasting frequencies found in text and in the world. In Section 3, we propose an explanation of reporting bias as a systematic distortion of reality. In Section 4, we look at how existing knowledge-extraction systems suffer from reporting bias and at attempts to correct for some forms of it. In Section 5, we suggest approaches for future work in knowledge extraction from text in light of our discussion.

2. MEASURING REPORTING BIAS

To demonstrate the reality of reporting bias and motivate our discussion in the next section, we will give several examples of the frequency of predications in text and in extractions that differ significantly from what we know about the world. Giving a full, accurate model of reporting bias or establishing how widespread the problem is would require the availability of real-world frequencies for all properties we're interested in learning from text. Instead, we simply demonstrate the existence of significant reporting bias for actions or events, outcomes, and properties.

We present textual frequencies based on the Google Web 1T n-gram data [3], which is derived from approximately a trillion words of Web text (circa 2006). We support this, where possible, with the number of times the Knext knowledge-extraction system learns a relevant claim about the world. Knext results are taken from a

[1] Knext is a system under development since before 2002 [25]. Here it is sufficient to consider it a more logically formalized set of interpretive rules than the later Stanford Dependencies [8, 9], leading to generic knowledge similar to that targeted by later systems such as NELL [5] (further details in Section 4).

Word	*Teraword*	*Knext*	*Word*	*Teraword*	*Knext*
spoke	11,577,917	244,458	hugged	610,040	10,378
laughed	3,904,519	169,347	blinked	390,692	20,624
murdered	2,843,529	11,284	was late	368,922	31,168
inhaled	984,613	4,412	exhaled	168,985	3,490
breathed	725,034	34,912	was punctual	5,045	511

Table 1: Frequencies from [3] and the number of times Knext learns that *A person may* $\langle x \rangle$, including appropriate arguments, e.g., *A person may hug a person*. For *murder*, more frequently encountered in the passive, we include *be murdered*.

knowledge base of six million unique factoids learned from a variety of corpora, including the Brown Corpus [20], the British National Corpus [2], Project Gutenberg e-books, Wikipedia, and the ICWSM 2009 weblog corpus [4]. This knowledge base is available to browse at http://cs.rochester.edu/research/knext/browse.

2.1 Events

One kind of knowledge that systems may try to extract from text is the typical frequency of an event or how characteristic an action is of a class of individuals, for generic claims such as *Generally people sleep* or *All or most people sleep daily*, while only *Some people play the fiddle*.

In Table 1, we see that murder is mentioned in text many more times than more quotidian actions like hugging or constant activities like breathing. We find people are late much more than they are punctual. While there are variations in the implied eventive frequencies between these two extraction methods, the gross level distortion of reality is clear.[2]

2.2 Outcomes

Another important kind of knowledge is the expected outcome of an action or event, e.g., *If a person drops a glass, it may break.* As this knowledge relies on larger patterns of predication, often involving more than one sentence, it is not easily measured on a large scale. A simple example, however, is while we know that for most races (whether foot races, political contests, etc.) the number of winners is less than or equal to the number of losers, we find more reports of a person winning a race than losing it: In the n-grams, 'won the race' occurs more than six times as often as 'lost the race'.

In Table 2, we see that, per mile travelled, a person is more likely to experience a crash on a motorcycle than in a car or in an airplane. However, motorcycle crashes are mentioned half as frequently in text as plane crashes despite their greater likelihood. We are not interested in replacing formal sources of data with these textual references, and commonsense knowledge doesn't require knowing how likely a crash is per mile traveled. Nonetheless, this discrepancy between reality and textual frequency may, as we believe, indicate a more pervasive distortion.

2.3 Properties

In the introduction, we used the example of how often we are told a person has a spleen vs having eyes. We are less interested in learning body parts from text than other knowledge, given the ease of acquiring this information from sources like WordNet [12] or simply enumerating it. We return to this example because we know that body parts are universally present in individuals (excepting abnormalities, accidents, surgical removal, and sexual differences). Therefore, when we see in Table 3 that Knext learns people have a head more than 1,000 times as often as it learns they have a pancreas, it makes it evident that we cannot take the frequency of reference to directly indicate the prevalence of a property.

Type	*Miles Travelled*	*Crashes*	*Miles/Crash*	*Teraword*
car	1,682,671 million	4,341,688	387,562	1,748,832
motorcycle	12,401 million	101,474	122,209	269,158
airplane	6,619 million	83	79,746,988	603,933

Table 2: *Miles Travelled*, *Crashes*, and *Miles/Crash* are for travel in the United States in 2006 [29]. Plane crashes are considered any event in which the plane was damaged. *Teraword* results are for the patterns 'car (crash|accident)', 'motorcycle (crash|accident)', and '(airplane|plane) (crash|accident)'

Body Part	*Teraword*	*Knext*	*Body Part*	*Teraword*	*Knext*
Head	18,907,427	1,332,154	Liver	246,937	10,474
Eye(s)	18,455,030	1,090,640	Kidney(s)	183,973	5,014
Arm(s)	6,345,039	458,018	Spleen	47,216	1,414
Ear(s)	3,543,711	230,367	Pancreas	24,230	1,140
Brain	3,277,326	260,863	Gallbladder	17,419	1,556

Table 3: N-gram results for '(his|her|my|your) *body part*' and the number of times Knext learned *A person may have a* $\langle$*body part*$\rangle$. Plurals are included when appropriate.

3. DISCUSSION

Reporting bias results from our responsibility as communicators to be maximally informative in what we convey to others who share our general world knowledge and to convey information in which they are likely to be interested.

The first of these imperatives was postulated by Paul Grice [17] as his *conversational maxim of quantity*. This states that communication should be as informative as necessary but no more, leaving unstated information that can be expected to be known or can be inferred from what is said using commonsense knowledge. Havasi et al. [18] previously related the difficulties of knowledge extraction to Gricean principles, writing "when communicating, people tend not to provide information which is obvious or extraneous. If someone says 'I bought groceries', he is unlikely to add that he used money to do so, unless the context made this face surprising or in question."

The second imperative – to be interesting – is less a linguistic principle than a psychological or social one. Some topics are inherently interesting, regardless of their prevalence, and we will tend to discuss these, biasing what information is available in text.

Note that even for properties that are quite common, if a person would not assume it, it's informative to mention it. For instance, if we describe a person we met, we may well say they have brown hair even though it's extremely common. However, we're even more likely to mention a person's hair color if it's unusual: While textual references to brown hair are more frequent than red (594,997 to 382,989), the latter's representation is quite disproportionate to its occurrence in the population.

3.1 Hypotheses

To elaborate and clarify this discussion, we offer these hypotheses about reporting bias, with corresponding examples:

1. The more expected the outcome of an action or event, the less likely people are to communicate it. E.g., we don't say 'I paid for the book and then I owned it' or 'A suicide bomber

[2] The variation is potentially due to differences in the text Knext has read vs the webpages the n-grams represent but also due to the more stringent requirements of a knowledge-extraction system. E.g., factoids about murder are discarded if they lack the complement (i.e., *A person may murder* vs *murder a person*).

blew himself yesterday. He died.' as these are assured consequences. However, we might say, 'I crashed my car. It was totalled.' as the degree of damage is not certain otherwise.

2. Reporting bias has a lot to do with the *values* we attach to things; we will give information even about an expected outcome if it is important to people. E.g., in a report of forest fires sweeping parts of California, we care about homes destroyed, and people killed or injured, but most care less about the number of chipmunks or deer killed. Further, the destruction of thousands of acres of forest will often matter and will be mentioned, as would the loss of members of a rare animal species.

3. People are unlikely to state usual properties of something as the primary intent of their utterance, e.g., 'a man with two legs' or 'a yellow pencil'. Rather, we state exceptional properties: 'a man with one leg', 'a blue pencil'. This bias also manifests in the lexicon: We talk about pregnant women, but there's no word for a woman who is not pregnant as it's the default assumption.

   While this means (as Havasi et al. [18] have claimed) we should not expect to acquire all common knowledge from text, even if commonsense isn't usually stated explicitly ('People use money to buy things'), it can appear as presuppositions, e.g., 'I forgot the money to buy groceries'. We discuss learning from such statements in Section 5.

4. Even unusual properties are unlikely to be mentioned if they're trivial. E.g., having a scratch on the left bicep may be as rare as pregnancy, but it usually matters too little to be reported.

5. Reporting bias will vary by literary genre. There will be considerable differences in the frequency of reporting events in an encyclopedia vs in fiction or even, e.g., among different newspapers. While sports pages will "over-report" sporting events compared to crimes, celebrity shenanigans, or business news, the National Inquirer or the Wall Street Journal might over-report other types of events.

6. There are fundamental kinds of lexical and world knowledge that are needed for understanding and inference that don't get stated in text, either because they are innate or because they are learned before language is acquired.

   We mean, e.g., physical objects can't be in different places at the same time; solid objects tend to persist (in shape, color and other properties) over time; if *A* causes *B* and *B* causes *C* then it's usually fair to say that *A* causes *C*; people do and say things for reasons – to get food or possessions or pleasure, to avoid suffering or loss, to make social connections, to provide or solicit information, etc.; you can't grab something that's out of reach; you can see things in daytime that are nearby and not occluded; people can't fly like birds or walk up or through walls; etc.

   There are also the lexical entailments and presuppositions that we learn as part of language and hardly ever say: 'above' and 'below', 'bigger' and 'smaller', 'contained in' and 'contains', 'good' and 'bad', etc., are incompatible; dying entails becoming dead; going somewhere entails a change in location; walking entails moving one's legs, etc.

4. PREVIOUS APPROACHES

In looking at how systems have dealt (or not dealt) with reporting bias, we want to contrast three lines of work: information extraction systems [7, 24], which learn explicitly stated material; knowledge extraction systems (e.g., [32]), which abstract individual instances to the general knowledge that's implicit in them; and systems that learn general rules implicit in a collection of specific extractions (e.g., [23, 31, 5]).

4.1 TextRunner

TextRunner [1] is a tool for extracting explicitly stated information as tuples of text fragments, representing verbal predicates and their arguments. After extraction, tuples are normalized (e.g., 'was originally developed by' to 'was developed by'). TextRunner's output includes both information about specific individuals and generic claims. Based on the number of distinct sentences from which a tuple was extracted, it is assigned a probability of being a correct instance of the relation.

The authors view the probabilities assigned to these claims not as representing the real-world frequency of an action or the likelihood the relation holds for an instance of a generic subject, but simply as the probability that the tuple is "a correct instance of the relation". It's not clear what this means for their "abstract tuples", which are 86% of the output on average, per relation, and include claims such as *(Einstein, derived, theory)* or *(executive, hired by, company)*. Is this a correct instance if Einstein at any point derived a theory? What if any executive was at some point hired by a company? Or is an abstract tuple only a correct instance of the relation if it's a generic claim – *Executives are (generally) hired by companies*?

4.2 Knext

Knext [25, 32] is a tool for extracting general world knowledge from large collections of text by syntactically parsing each sentence with a Treebank-trained parser (e.g., [6]) and applying interpretive rules for computing logical forms in a bottom-up sweep, abstracting those that serve as stand-alone propositions. The results are quantificationally underspecified Episodic Logic formulas, which are verbalized in English as possibilistic claims, e.g., *Persons may want to be rid of a dictator*. Knext treats all discovered formulas as possible general world knowledge. In an evaluation of 480 propositions, Van Durme et al. [32] observed that propositions found at least twice were judged more acceptable than those extracted only once. However, as the support increased above this point, the average assessment stayed roughly the same. That is, frequency of extraction was not found to be a reliable indication of quality.

Later work [13] has sharpened Knext output into explicitly quantified, partially disambiguated axioms. This used the pointwise mutual information between subject terms and what's predicated of them as one of the factors in determining appropriate quantifier strength. However, this association is overruled by semantic patterns, e.g., having a body part is (near-)universally true for a class of individuals even if – as with people's spleens – it is rarely mentioned. For other predications, such as having a particular kind of possession, these sharpened axioms are subject to the distortion of reporting bias.

4.3 Textual Frequency and Probability

TextRunner's probabilities use the model of Downey et al. [11], which is based on the belief that "An extraction that is obtained from multiple, distinct documents is more likely to be a bona fide extraction than one obtained only once. Because the documents that 'support' the extraction are, by and large, independently authored, our confidence in an extraction increases dramatically with the number of supporting documents.” This may be true for specific facts, e.g., *Einstein was born in 1879*, but the evaluation of Knext extractions suggests the same does not hold for the generic claims we seek: For general world knowledge, additional sources do not correlate with better quality.

However, it is possible that more of a correlation would be seen when looking only at distinct corpora (or, in a heterogenous corpus like a collection of Project Gutenberg e-books, distinct sources within the corpus). A stronger correlation might also be seen if we count distinct constructions rather than whole sentences: The occurrence of fixed phrases, such as titles or idioms, can lead to bad claims. E.g., the film ‘True Lies’ – often misparsed as a common noun phrase – leads Knext to learn *Lies may be true*, just as the idiom ‘when pigs fly’ gives us *Pigs may fly*. While biasing the acceptability of a claim on the number of distinct constructions from which it is learned would increase quality, it wouldn’t solve the general problem of reporting bias we’ve presented.

4.4 Learning Implicit Rules from Extracted Facts

A line of work at Oregon State University [27, 10] learns domain-particular rules based on specific facts extracted from text. They address a subproblem of the general reporting-bias phenomenon, namely the conditional bias of our Hypothesis 3. If attribute $A(x) = a$ of some entity is reported, and $A(x) = a$ tends to imply $B(x) = b$, then $B(x) = b$ tends not to be reported. (E.g., if someone is stated to be a Canadian citizen, then we are less likely to also state that they were born in Canada.) But if, in fact, $B(x) = b'$, then we are likely to say so. (E.g., we *would* say ‘an Egyptian-born Canadian’.)

Raghavan and Mooney [23] learn commonsense knowledge in the form of probabilistic first-order rules from the incomplete, noisy output of an information-extraction system. Their rules have a body containing relations that are often stated explicitly, while the head uses a relation that is mentioned less often as it’s easily inferred. They produce rules like

> hasBirthPlace(x, y) $\wedge$ person(x) $\wedge$ nationState(y)
> $\rightarrow$ hasCitizenship(x, y)

An interesting aspect of their approach is the use of WordNet similarity to weight rules, based on the idea that more accurate rules usually have predicates that are closely related in meaning.

5. ADDRESSING REPORTING BIAS

We’ve shown that reporting bias’s distortion of real-world frequency in text makes it doubtful that we can interpret the number of textual references or explicit statements supporting a general claim as directly conveying real-world prevalence or reliability. While there seems to be no silver bullet, there are some approaches to learn what normally holds in the world. For instance, we can ignore frequency and focus on more informative extraction targets:

1. *Disconfirmed expectations*. Gordon & Schubert [14] learned commonsense inference rules from constructions that indicate a speaker’s expectation about the world was not met, e.g.,

   > ‘Sally crashed her car into a tree but wasn’t hurt.’
   > → *If a person crashes her car, she may be hurt.*

2. *Implicit denials.* Explicit statements, pragmatically required to be informative, contain implicit denials that what they’re saying is usually the case. However, these vary in how easily they can be transformed into the implicit claims, e.g.,

   > ‘The tree had no branches.’
   > → *Trees usually have branches.*

   > ‘Molly handed me a blue pencil.’
   > → *Probably pencils are not usually blue.*

   We can look for more explicit denials, but they defy the maxim of quantity by making commonsense explicit and thus are rare. E.g., we are unlikely to say, ‘She handed me a pencil, but it wasn’t a normal yellow one; it was blue!’

3. *Presupposed numbers and frequencies*. Some constructions show a presupposition – a belief the speaker expects others to share – about how many of a thing are normal, e.g.,

   > ‘Both my legs hurt.’
   > → *A person normally has two legs.*

   Other patterns presuppose what is considered exceptional, as in the patterns of event frequency evaluated by Gordon & Schubert [15]:

   > ‘I hadn’t slept in days.’
   > → *A person normally sleeps at least daily.*

   (These claims about event frequencies are implicitly conditioned on whether the agent does the action at all. E.g., *If a person writes a book at all, he probably does so every few years.*)

Another possibility is to use a hybrid approach to knowledge extraction, along the lines of [26] or [19]. For instance, we might combine text mining with a crowdsourced rating [16] or filtering stage to assign an approximate real-world frequency to the knowledge found most frequently in text.

Finally, we suggest using textual frequencies for the problems where a view of the world skewed by what is unusual and interesting is actually helpful. One such problem is selecting appropriate axioms for forward inference. Even for a small knowledge base, it is infeasible to generate all possible inferences. Rather, we want to focus on those most likely to be important. If we are told ‘John is a person’, we don’t want to reason about very probable but trivial properties, such as his having skin cells. Rather, we want to reason that he probably has a job of some kind, that he lives somewhere, and so on – facts more likely to be important to further reasoning.

6. CONCLUSIONS

We have argued that researchers need to be aware that frequency of occurrence of particular types of events or relations in text can represent significant distortions of real-world frequencies and that much of our general knowledge is never alluded to in natural discourse. We provided a brief pragmatic argument for why reporting bias exists, which led to suggestions on how we might, partially, work around it.

Our examples and discussion are meant to provoke further study. If reporting bias is not a real problem for knowledge acquisition, it remains for the community to show this to be the case. Otherwise, more work is called for to determine if, and how, we can correct for it. At worst, reporting bias may prove an upper bound on the extent to which human knowledge can be learned from text and may provoke further work on hybrid approaches to knowledge acquisition.

Acknowledgments

This work was supported by NSF IIS-0916599 and a subcontract to ONR STTR contract N00014-10-M-0297. We are grateful to Lenhart Schubert for helpful comments.

7. REFERENCES

[1] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open Information Extraction from the Web. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2007.

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

[3] T. Brants and A. Franz. Web 1T 5-gram Version 1. Distributed by the Linguistic Data Consortium, 2006.

[4] K. Burton, A. Java, and I. Soboroff. The ICWSM 2009 Spinn3r dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM)*, 2009.

[5] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI)*, 2010.

[6] E. Charniak. A maximum-entropy-inspired parser. In *Proceedings of the First Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 132–9, 2000.

[7] J. Cowie and W. Lehnert. Information extraction. *Communications of the Association for Computing Machinery*, 39:80–91, 1996.

[8] M.-C. de Marneffe, B. Maccartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of the International Conference on Language Resources and Evaluation*, 2006.

[9] M.-C. de Marneffe Christopher D. Manning. The Stanford typed dependencies representation. In *Proceedings of the COLING workshop on Cross-Framework and Cross-Domain Parser Evaluation*, pages 1–8, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[10] J. R. Doppa, M. NasrEsfahani, M. S. Sorower, T. G. Dietterich, X. Fern, and P. Tadepalli. Towards learning rules from natural texts. In *Proceedings of the NAACL Workshop on Formalisms and Methodology for Learning by Reading*, pages 70–77, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[11] D. Downey, O. Etzioni, and S. Soderland. A probabilistic model of redundancy in information extraction. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2005.

[12] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[13] J. Gordon and L. K. Schubert. Quantificational sharpening of commonsense knowledge. In *Proceedings of the AAAI Fall Symposium on Commonsense Knowledge*, 2010.

[14] J. Gordon and L. K. Schubert. Discovering commonsense entailment rules implicit in sentences. In *Proceedings of the EMNLP Workshop on Textual Entailment (TextInfer)*, 2011.

[15] J. Gordon and L. K. Schubert. Using textual patterns to learn expected event frequencies. In *Proceedings of the NAACL Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction (AKBC-WEKEX)*, 2012.

[16] J. Gordon, B. Van Durme, and L. K. Schubert. Evaluation of commonsense knowledge with Mechanical Turk. In *Proceedings of the NAACL 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010.

[17] H. P. Grice. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, San Diego, CA, 1975.

[18] C. Havasi, R. Speer, and J. Alonso. ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In *Proceedings of Recent Advances in Natural Language Processing*, 2007.

[19] R. Hoffman, S. Amershi, K. Patel, F. Wu, J. Fogarty, and D. S. Weld. Amplifying community content creation with mixed-initiative information extraction. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2009.

[20] H. Kučera and W. N. Francis. *Computational Analysis of Present-Day American English*. Brown University Press, 1967.

[21] D. B. Lenat. Cyc: A Large-scale Investment in Knowledge Infrastructure. *Communications of the Association for Computing Machinery*, 38(11):33–48, 1995.

[22] S. Moody. The brain behind Cyc. *The Austin Chronicle*, 1999. http://www.austinchronicle.com/screens/1999-12-24/75252/.

[23] S. Raghavan and R. J. Mooney. Online inference-rule learning from natural-language extractions. July 2013.

[24] S. Sarawagi. Information extraction. *Foundations and Trends in Databases*, 1:261–377, 2008.

[25] L. K. Schubert. Can we derive general world knowledge from texts? In *Proceedings of the Second International Conference on Human Language Technology Research (HLT)*, 2002.

[26] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast – but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, 2008.

[27] S. Sorower, T. G. Dietterich, J. R. Doppa, W. Orr, P. Tadepalli, and X. Fern. Inverting Grice's maxims to learn rules from natural language extractions. In *Proceedings of Neural Information Processing Systems (NIPS)*, pages 1053–61, 2011.

[28] D. G. Stork. The Open Mind Initiative. *IEEE Expert Systems and Their Applications*, pages 16–20, May/June 1999.

[29] U.S. Department of Transportation. National transportation statistics. http://www.bts.gov/publications/national_transportation_statistics, October 2009.

[30] B. Van Durme. *Extracting Implicit Knowledge from Text*. PhD thesis, University of Rochester, 2010.

[31] B. Van Durme, P. Michalak, and L. K. Schubert. Deriving generalized knowledge from corpora using WordNet abstraction. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2009.

[32] B. Van Durme and L. K. Schubert. Open knowledge extraction through compositional language processing. In *Proceedings of the Symposium on Semantics in Text Processing (STEP)*, 2008.