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 despite this distortion.

Categories and Subject Descriptors

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

Keywords

Knowledge extraction; text frequency; reporting bias

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 a large amount 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 suitable knowledge (as in the Cyc project [22]) or to solicit it directly from crowds online (as in the Open Mind Initiative [30]), the dominant approach is to mine knowledge from the extensive text available in electronic form.

A 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 the atypical or claims that are simply not true. For instance, the Knext [27] system for knowledge extraction from text (described in Section 4) learns both *The Earth may revolve around the Sun* and *The Sun may revolve around the Earth*. Mistaken claims like the latter may indicate a failure to correctly learn from text (e.g., if a source said 'It is *not* the case that the Earth revolves around the Sun'), or it may result from reading an inaccurate or fantastical text.

To identify good claims, it is typical to take an inductive view, with textual references serving as evidence: The more often we read something, the more likely it is to reflect what is true in the real world. This is intuitively reasonable, and, over a large collection of texts, Knext learns the heliocentric claim 327 times, while the geocentric claim is only learned 126 times. However, the frequency with which situations of a certain type are described in text does not always correspond to their relative likelihood in the world, or even the subjective frequency captured in human beliefs. For instance, from the same texts, Knext learns almost a million times that *A person may have eyes*, but fewer than 1,600 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*.[1]

For knowledge extraction (KE), we are interested in reporting bias as it relates to the frequency with which events or actions occur, the likelihood of specific outcomes, and the prevalence 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 Douglas 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." [23]

While the focus of this paper is on how reporting bias affects the acquisition of general knowledge, many of the phenomena we discuss also apply to factual information extraction (IE). E.g., frequently reading claims that Barack Obama was born in Kenya does not make it a reliable extraction. However, for a factual IE system, other extraction properties may be more salient than textual frequency. For instance, the great frequency of statements that George Bush is the president of the United States should not lead us to believe this is currently true, given the greater recency of sentences indicating Obama is president. The trustworthiness of text sources can also be of greater importance for factual IE than for general knowledge extraction, which can abstract claims even from realistic fiction.

2. MEASURING REPORTING BIAS

To demonstrate the reality of reporting bias and motivate our discussion in the next section, we will give several examples where the frequencies of textual references and extractions differ significantly from what we know to be the case in 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 across

1 This article is an expansion on points raised in the Ph.D. thesis of the second author [32].

AKBC '13, October 27–28, 2013, San Francisco, CA, USA.

ACM 978-1-4503-2411-3/13/10 … $15.00.
http://dx.doi.org/10.1145/2509558.2509563.

Table 1: N-gram frequencies for (*his*|*her*|*my*|*your*) ⟨*body part*⟩ and the number of times Knext learns *A* ⟨*body part*⟩ *may pertain to a person*. Plurals are included when appropriate.

Body Part	*Teraword*	*Knext*	*Body Part*	*Teraword*	*Knext*
Head	18,907,427	1,004,300	Liver	246,937	9,452
Eye(s)	18,455,030	934,721	Kidney(s)	183,973	3,289
Arm(s)	6,345,039	399,120	Spleen	47,216	1,568
Ear(s)	3,543,711	309,708	Pancreas	24,230	1,186
Brain	3,277,326	144,511	Gallbladder	17,419	991

Table 2: N-gram frequencies for various verbal events and the number of times Knext learns that *A person may* ⟨*x*⟩, including appropriate arguments, e.g., *A person may hug a person*.

Word	*Teraword*	*Knext*	*Word*	*Teraword*	*Knext*
Spoke	11,577,917	372,042	Hugged	610,040	11,453
Laughed	3,904,519	179,395	Blinked	390,692	21,973
Murdered	2,843,529	16,890	Was late	368,922	31,168
Inhaled	984,613	5,617	Exhaled	168,985	4,052
Breathed	725,034	41,215	Was on time	23,997	14

the range of types of properties that we are 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 Knext learns a relevant claim about the world. These results are taken from a knowledge base of 73 million unique factoids learned from sources including the Brown Corpus [21], the British National Corpus [2], Gigaword [24], Project Gutenberg books, Wikipedia, and the ICWSM 2009 weblog corpus [4]. The full knowledge base is available to browse at http://cs.rochester.edu/research/knext/browse.

In the introduction, we used the example of how often we are told a person has spleen vs having eyes. In Table 1, we see the significant variation with which body parts are mentioned in writing, though they are near universally present in individuals. While this type of knowledge is readily available from manually created sources such as WordNet [13] or Cyc [22], the fact that even such simple extractions exhibit significant reporting bias bodes ill for the long tail of more subtle knowledge that we are less likely to be able to enumerate.

For instance, KE systems may try to learn from text the typical frequency of an event or how characteristic an action is of a class of individuals, to produce generic claims such as *Generally people sleep* or *Most people sleep daily*, while only *Some people play the viola*. However, in Table 2, we see that murder is mentioned in text many more times than more quotidian actions like hugging or constant activities like breathing, and we find people are late much more than they are on time. The Knext extraction frequencies can be seen as a further distortion of the textual frequencies, due, at least in part, to the filtering of potential claims. For instance, factoids about murder are automatically discarded if they lack the complement (i.e., you need to murder *someone*).

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. However, in Table 3 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, in text motorcycle crashes are only mentioned half as frequently as plane crashes.

Table 3: *Miles Travelled*, *Crashes*, and *Miles/Crash* are for travel in the United States in 2006 [31]. A plane crash is 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*).

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

For a simpler example, 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, yet we find far 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* (66,011 vs 10,430). The number of matches for (*participated in*|*ran in*|*took part in*|*entered*) *the race*, which lack the stigma of "losing", is still quite low (22,512). Even for the Academy Awards, where "it's an honor just to be nominated", people are much more likely to write about a win than a nomination. We find *won the academy award* 15,098 times vs 4,551 for *nominated for the academy award* (and the same is true for a number of variations, such as *academy award winner* and *academy award winning*).

3. DISCUSSION

We believe these discrepancies between reality and textual frequency indicate a pervasive distortion. Reporting bias results from our responsibility as communicators to be maximally informative in what we convey to other people, 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 [18] 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. [19] previously knowledge acquisition from text to Gricean principles, noting that "people tend not to provide information which is obvious or extraneous" and, therefore, "it is difficult to automatically extract common-sense statements from text, and the results tend to be unreliable". The second imperative – to be interesting – is less a linguistic principle than a psychological or social one: Some topics are intriniscally interesting to people, regardless of their prevalence, and we will tend to discuss these, biasing what information is available in text.

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

1. *The more expected something, the less likely people are to convey it as the primary intent of an utterance.*

People are unlikely to tell you about 'the man with two legs' or 'a yellow pencil'. Rather, we state exceptional properties: 'a man with one leg', 'a blue pencil'. Similarly, we don't say 'I paid for the book and then I owned it' or 'A suicide bomber blew himself yesterday. He died.' as these are assured consequences. We might, however, say, 'I crashed my car. It was totalled.' as the degree of damage is not certain otherwise. While expected information is unlikely to be the *primary purpose* of an utterance, it can appear in presuppositions; see Section 5.

2. *The more value people attach to something, the more likely they are to give information about it, even if the information is unsurprising.*

For instance, 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. If we describe a person we met, we may well say he has brown hair even though this extremely common. However, we are 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 in the Google n-grams), the latter's representation is quite disproportionate to its occurrence in the population.

3. *Conversely, even unusual facts are unlikely to be mentioned if they are trivial.*

E.g., having a scratch on the left bicep may be less common than an interesting, important property like a woman being pregnant, but it usually matters too little to be reported.

4. *Reporting bias varies 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.

5. *There are fundamental kinds of lexical and world knowledge that are needed for understanding and inference that don't get stated in text.*

This can be because they are innate or are learned before language is acquired, by physical interaction in the world. 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 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 are 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, 26], which learn explicitly stated material; knowledge extraction systems (e.g., [34]), 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., [25, 33, 5]). We only provide a few examples; for a more thorough overview, see [32].

TextRunner

TextRunner [1] is a tool for extracting explicitly stated information as tuples of normalized text fragments, representing verbal predicates and their arguments. 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. TextRunner's 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 is a good generic statement, e.g., *Executives are (generally) hired by companies*?

Knext

Knext [27, 34], under development since before 2002, 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 compositionally applying interpretive rules to compute logical forms in a bottom-up sweep, abstracting those that serve as stand-alone propositions. (It's 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 systems subsequent to Knext, such as NELL [5].) 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. [34] 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 [14] has sharpened Knext output into explicitly quantified, partially disambiguated axioms. This uses 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 classes of predication, such as having a possession, these sharpened axioms are subject to the distortion of reporting bias.

Urns

TextRunner's probabilities use the Urns model of Downey et al. [12, 11], which is based on the belief that an extraction is more likely to be true if it is obtained from multiple documents, adjusting for how often a type of reference occurs. E.g., Urns should assign a lower probability to 'countries such as Washington' (31 hits on the Web) than it does to 'throwable objects such as bean bags' (3 hits) given the far greater number of extractions for countries than for throwable objects (example due to Doug Downey). However, Urns is meant to establish the truth of ground facts (e.g., *Einstein was born in 1879*), not the probability of a generic claim applying (e.g., *People eat food*). Indeed, a great deal of the commonsense knowledge we want to learn is only discovered a handful of times, even over Web-scale text, while Urns requires a fairly large sample size for each relation.

Learning Rules from Extracted Facts

A line of work at Oregon State University [29, 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 1. 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'.)

Along similar lines, Raghavan and Mooney [25] 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 $\text{hasBirthPlace}(x, y) \wedge \text{person}(x) \wedge \text{nationState}(y) \rightarrow \text{hasCitizenship}(x, y)$. 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 focus extraction on more informative constructions:

1 *Presuppositions*. Commonsense knowledge that is rarely stated explicitly can nonetheless appear in sentences as presuppositions – beliefs the speaker expects others to share:

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

> 'I forgot the money to buy groceries.'
> → *A person may use money to buy things.*

2 *Disconfirmed expectations*. Gordon & Schubert [15] 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.*

> 'I dropped the glass, but it didn't break.'
> → *If a person drops a glass, it often will break.*

Other sentences suggest that an action or event has not taken place with the normal temporal frequency [16]:

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

(These claims 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.*)

3 *Implicit denials.* Explicit statements, pragmatically required to be informative, contain implicit denials that what they're saying is usually the case, e.g.,

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

However, these vary in how easily they can be transformed into general claims, e.g.,

> 'Molly handed me a blue pencil.'
> → *Probably pencils are not always blue.*

4 *Reference to individuals*. Expected properties can be expressed when identifying a particular individual, e.g.,

> '…the man I met yesterday.'
> → *A person may meet a man.*

Claims frequently learned from such constructions may be more usual than those learned from more explicit assertions, though there are still many more references to a 'plane that crashed' than a 'plane that landed'.

More correlation might be seen between frequency and extraction quality if we only count the frequency of distinct textual references. E.g., repeated mentions of the film *True Lies*, misparsed as a common noun phrase, lead Knext to learn *Lies may be true*. Even if text is analyzed correctly for its surface meaning, it can lead to bad knowledge, e.g., the idiom 'when pigs fly' gives us *Pigs may fly*. A related problem is frequently repeated text, such as song lyrics on the Web. To account for this textual bias – exact repetition – we might give more weight to knowledge learned from different extraction methods or just from distinct constructions.

Another possibility is to use a hybrid approach to knowledge extraction, along the lines of [28] or [20]. For instance, we might combine text mining with a crowdsourced rating [17] or filtering stage to assign an approximate real-world frequency to the knowledge found most frequently in text. Work in the emerging "grounded language" movement may also be important. If one were to say 'John entered the room', they are unlikely to follow it up with 'He blinked. He breathed.' However, many mundane actions and activities might be recognized, e.g., by sampling video and be incorporated into our knowledge.

It is also important to recognize that for some problems, frequencies for the distorted world described in text are more useful than real-world frequencies. For instance, a parser is concerned with how frequently 'cat' is the subject of 'meow', rather than how frequently cats actually meow. With the bias for the interesting or unusual, textual frequencies may also be useful for guiding inference for conclusions that are most likely to be important or useful: If we are told 'John is a person', we don't want to reason that he has skin cells (although this is certainly true) but rather that he probably has a job of some kind, that he lives somewhere, etc.

6. CONCLUSIONS

We have argued that researchers need to be aware that the 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 ONR STTR N00014-10-M-0297. We are grateful to Lenhart Schubert, Peter Clark, Doug Downey, and our anonymous reviewers for their feedback.

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* (*IJCAI*), 2007.

[2] BNC Consortium. The British National Corpus, version 2. Distributed 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* (*LREC*), 2006.

[9] M. C. de Marneffe and C. 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. Etizioni, and S. Soderland. Analysis of a probabilistic model of redundancy in unsupervised information extraction. *Artificial Intelligence*, 174(11):726–48, July 2010.

[12] 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* (*IJCAI*), 2005.

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

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

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

[16] 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*), June 2012.

[17] 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.

[18] 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.

[19] 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* (*RANLP*), 2007.

[20] 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.

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

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

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

[24] C. Napoles, M. Gormley, and B. Van Durme. Annotated Gigaword. In *Proceedings of the NAACL Workshop on Automatic Knowledge Base Construction and Web-Scale Knowledge Extraction* (*AKBC-WEKEX*), June 2012.

[25] S. Raghavan and R. J. Mooney. Online inference-rule learning from natural-language extractions. In *Proceedings of the AAAI Workshop on Statistical Relational AI* (*StaRAI-13*), July 2013.

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

[27] 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.

[28] 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.

[29] 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.

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

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

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

[33] 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.

[34] 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.