Learning from the Web: Extracting General World Knowledge from Noisy Text

Jonathan Gordon
Department 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

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

Abstract

The quality and nature of knowledge that can be found by an automated knowledge-extraction system depends on its inputs. For systems that learn by reading text, the Web offers a breadth of topics and currency, but it also presents the problems of dealing with casual, unedited writing, non-textual inputs, and the mingling of languages. The results of extraction using the KNEXT system on two Web corpora – Wikipedia and a collection of weblog entries – indicate that, with automatic filtering of the output, even ungrammatical writing on arbitrary topics can yield an extensive knowledge base, which human judges find to be of good quality, with propositions receiving an average score across both corpora of 2.34 (where the range is 1 to 5 and lower is better) versus 3.00 for unfiltered output from the same sources.

Introduction

The creation of intelligent artifacts, capable of human-level reasoning, requires considerable knowledge. This *knowledge acquisition bottleneck* is apparent in efforts to improve parsing, question-answering, and other difficult ("AI-complete") problems. *Information extraction* efforts, *e. g.* Banko *et al.* (2007), have focused on learning facts about specific entities, such as that Alan Turing died in 1954 or that the capital of Bahrain is Manama. Knowledge bases of such facts are quite useful, but getting to human-level AI seems to depend less on this specific knowledge than it does on the most basic world knowledge – our commonsense understanding of the world.

For instance, to correctly choose the most likely syntactic parse of the sentence "I saw a robin with my binoculars", it would help to have the knowledge that a robin is a bird and that birds (and all non-human animals) are unlikely to *have* binoculars (or other man-made tools), but they are often seen *through* binoculars (and other artifacts such as rifle sights, cameras, *etc.*). Thus an intelligent parser would attach "with my binoculars" to "saw" and not to "a robin".[1]

[1]The problem of improving parsing using shallow world knowledge like that discussed in this paper requires a bootstrapping cycle since KNEXT itself depends on syntactic parsing. This was argued for by Schubert (2009), and the potential effectiveness of such an approach was demonstrated by Clark and Harrison (2009).

This kind of general world knowledge is rarely stated directly, but it is implicit in written language. To discover propositions like these, Schubert (2002) created KNEXT (Knowledge Extraction from Text). KNEXT is an *open knowledge extraction* system, meaning that rather than seeking out a set of relations about specific individuals or targeting a restricted class of general relations, it looks for all world knowledge that can be learned from a text, abstracting from individuals to generic *kinds* whenever possible and storing the resulting knowledge as symbolic logical formulas, not tuples of strings or graphs of relations.

These formulas are automatically translated back into approximate English representations, giving *factoids*, such as 'A CHILD MAY WRITE A POEM', 'A PERSON MAY SAY SOMETHING TO A GROUP', and even epistemic ones like 'A PERSON MAY UNDERSTAND AN ALLURE OF PART OF A BOOK' or 'A PERSON MAY SEEM TO FIGURE_OUT A THING FROM A WEBSITE'. Note that rather than stating what necessarily holds, the factoids express what is possible in the world.

KNEXT operates by applying compositional semantic interpretation rules to extract knowledge from syntactic parses of English text.[2] While it was originally (Schubert and Tong 2003) used on the hand-parsed Brown Corpus (Kučera and Francis 1967), KNEXT was extended to use third-party statistical parsers, allowing the use of other corpora. The use of webtext for knowledge extraction presents new challenges, which were reported by Gordon, Van Durme, and Schubert (2009).

We greatly increase the data set of that work and introduce a method for filtering the resulting propositions to find a core set of high-quality knowledge. We offer an assessment of the quality of knowledge that can be learned from unstructured, unedited weblog text and from the more edited, knowledge-oriented writing of Wikipedia – with and without such filtering – and consider whether weblogs could be a worthwhile source for knowledge mining compared with Wikipedia. The contributions of the paper concern the following questions:

1. Does the volume of extracted general factoids grow indefinitely as more and more weblog sentences are processed

[2]A core version of KNEXT is being released at http://www.cs.rochester.edu/research/knext

(up to hundreds of millions), and similarly as Wikipedia sentences are processed?

2. Does factoid quality depend significantly on the two types of sources?
3. Can extraction quality be significantly improved using a collection of filtering techniques, such as removal of factoids that fail logical-form parsing, violate verb arity constraints, or contain many unlexicalized word stems?
4. To what extent do weblog-derived factoids cover Wikipedia-derived factoids and *vice versa*?

We show that the answers to (1) and (3) are positive, while the answers to (2) and (4) are "less than might be expected". It is expected that these results apply to other efforts to learn general world knowledge, such as Boeing's DART (Clark and Harrison 2009).

Web Corpora

The ease of publishing online has created an instantly-available, up-to-date, and increasingly comprehensive store of human knowledge, opinion, and experience. These same features which attract human readership motivate the Web as a resource for automated knowledge acquisition.

Traditional corpora will usually possess certain domain biases, which are undesirable for knowledge extraction: Project Gutenberg's collection of public-domain books may contain little knowledge about cellphones but plenty about telegrams; US newswire circa 1999 will have exhaustive knowledge about impeaching a president, but it probably has little that can be learned about dreaming or owning a cat.

Rather than worry about constructing ever-larger balanced collections of text to use with knowledge-acquisition systems like KNEXT, we are interested in discovering whether the vast amount of ungrammatically written ("noisy"), unedited, unfocused writing that can be found on the Web can prove an adequate substitute for what we might learn from other sources. Our question is thus: can we find a subset of usable knowledge amid the typos?

Weblogs

In 2009, the Third International AAAI Conference on Weblogs and Social Media released a large data set (Burton, Java, and Soboroff 2009) of 62 million postings to weblogs (and other sites that use syndication feeds, including some news or shopping sites), totalling 203 gigabytes. The data set was collected by Spinn3r.com between August and October 2008.

Much of the content in this data set is not in English or does not constitute *writing* – rather, it is the result of people posting pictures, videos, snippets of code, or spam text. Even the English writing is rarely straightforward, consisting of song lyrics, sentence fragments strewn with emoticons, or unpunctuated train-of-thought. Another complication is that since the data set originates from RSS and Atom feeds, many of the entries are only snippets of longer posts, and the truncation can occur mid-sentence.

To make this data set more readily parsable, we stripped the HTML tags (marking paragraphs, formatting text, embedding media, *etc.*), eliding text inside tags that indicate content we're unlikely to handle correctly, such as ⟨table⟩s, ⟨code⟩ fragments, or ⟨pre⟩formatted text, which is likely to be ASCII art or source code. Although we remove text that is not identified in the data set's XML as being English, much foreign writing is still included. The parser blithely treats such sentences as English, leading KNEXT to produce nonsensical factoids, which are subsequently filtered out.

Performing some simple text replacements before parsing can enhance the later extraction. For instance, "this website"[3] is substituted for URLs and "this email address" for all email addresses. Although these replacements oversimplify the way such addresses can be used in writing, they allow us to make some sense of them. *E.g.*, from the sentence "*some-site*.com posted an interesting link" we learn that 'A WEBSITE MAY POST A LINK'.

These substitutions can be considered a medium-specific augmentation of the usual KNEXT abstraction, which uses hand-constructed gazetteers to turn named entities into types, *e.g.*, *Bundestag* to 'A LEGISLATURE' or *John von Neumann* to 'A SCIENTIST'. Additionally, we applied a set of simple substitutions to correct common misspellings and accommodate the casual mode of writing often found online so that, for instance, "u r" is changed to "you are", making a correct parse more likely.

After this preprocessing, the weblog data set was reduced to 245,361,917 recognized sentences (26 gigabytes) – just 12% of the original data set. Such heavy pre-filtering of the weblog text reflects an interest in precision over recall, a typical preference when using web-scale data.

Wikipedia

Wikipedia is perhaps the most interesting source for knowledge-extraction efforts, both because of the great diversity of topics it describes and because of its mix of writing styles, ranging from high-profile articles with much-edited language to article stubs consisting of one person's random scribblings, waiting to be deleted. As such, it represents a middle ground between the formality of many traditional corpora and the free-for-all nature of weblogs.

Wikipedia articles are written for the express purpose of conveying accurate information about the world, not opinions, anecdotes, *etc.* This might seem to make Wikipedia the obvious best choice for knowledge extraction, but it is a resource for facts stated *explicitly* while KNEXT targets the general world knowledge that is found *implicitly* in writing. For instance, for the Wikipedia sentence "The emperor was succeeded by his son, Akihito", what we seek to (and do) learn is that 'AN EMPEROR MAY BE SUCCEED-ED BY A SON' and 'A MALE MAY HAVE A SON' – not the specific information about Emperor Shōwa and his son. Thus, having been written as a repository of information, which most weblogs are not, is not a clear advantage for Wikipedia as a resource for extracting background knowledge. If weblogs (and similar unstructured, untargeted text, *e.g.*, forum posts)

[3] Given KNEXT's extraction, this is equivalent to "a website".

	Sentences	*Words*	*Raw Factoids*	*Unique Factoids*	*Raw/100 words*	*Uniq./100 words*
Weblogs	95,296,872	2,004,492,555	202,282,757	67,632,550	10.1	3.4
Wikipedia	53,971,864	909,756,011	104,287,529	53,945,110	11.5	5.9
NY Times	39,433,116	773,074,059	124,956,881	43,939,886	16.2	5.7
Brown	51,763	1,026,595	132,314	109,443	12.9	10.7

Table 1: The numbers of factoids extracted from Web and traditional corpora. Sentence counts are the number of sentences parsed and then used for knowledge extraction, which in the case of the weblogs is smaller than the total available corpus.

can be of the same utility, they would be a more attractive resource since there is more of such text than Wikipedia article text.

For these experiments, we used a complete snapshot of English Wikipedia[4] (as of July 2, 2009) encoded as XML. It was stripped of Wiki markup, links, and figures using a tool by Antonio Fuschetto of the University of Pisa.[5]

Extraction and Comparisons

We ran a parser (Charniak 2000) over the Wikipedia snapshot and a sample of the weblog corpus approximately twice as large and then ran KNEXT on the resulting parse trees. The number of raw factoids produced (that is, the number before any filtering) can be seen in Table 1 along with the number of factoids produced per 100 words – the extraction density. For comparison, the same results are shown for two more traditional corpora: the Brown corpus and the New York Times portion of Gigaword (Graff *et al.* 2007). KNEXT produces more raw factoids from the same amount of weblog text than it does from Wikipedia, while there is no clear split in the extraction rates between the Web corpora and the traditional corpora. The factors that affect this, such as the number of modifiers used, differ in each corpus.

The Brown Corpus shows the highest extraction rate for *unique* factoids, as it is by design topically varied and non-repetitive (and its hand-crafted parses yield more clause-based, as opposed to modifier-based, factoids). In a potential reflection of the content, we find that weblogs yield fewer unique factoids for the same amount of text – we are more likely to learn the same things repeatedly from weblogs than from Wikipedia. However, as shown in Figure 1, as the number of raw factoids generated increases, the number of unique factoids generated only falls off slightly.

This means that there is a fairly consistent benefit to reading more text from each source. Since the amount of weblog text (and other casual, undirected writing on the Web) in existence is vast and continues to grow, a knowledge extraction system like KNEXT can continue to learn more about the world from the Web almost indefinitely: Any significant fall-off in results won't occur until after many hundreds of millions of sentences are read. While Wikipedia is also growing, its standards for worthy topics and for providing sources imply that text is added more slowly; even as new writing is added, other parts are being deleted.

[4] http://en.wikipedia.org/wiki/Wikipedia_database

[5] http://medialab.di.unipi.it/wiki/Wikipedia_Extractor

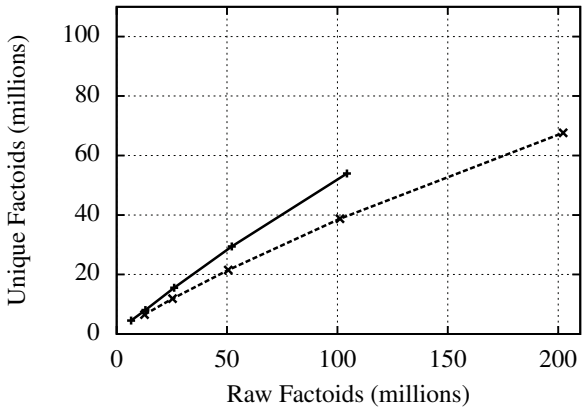

Figure 1: The growth of unique factoids found in each source as more raw factoids are generated. The dashed line is for the weblog corpus; the solid line for Wikipedia.

Looking forward to the use of such data, we might ask[6] whether a knowledge base that continues to grow indefinitely is a good thing. The answer is a qualified yes: As we continue to acquire more knowledge, the knowledge we haven't seen before is more likely to be about specific individuals or esoteric attributes. Thus there is a declining utility to learning more. However, when we seek to abstract from specific knowledge to more general truths that are unlikely to be stated in text, as in the work of Van Durme, Michalak, and Schubert (2009) and in future work, processing large volumes of text may result in better generalizations.

KNEXT generally learns a different set of factoids from weblogs than it does from Wikipedia. Only 5,226,089 unique factoids are found in exactly the same form in the two corpora. This means that just 7% of what we learn from the weblogs can also be found in Wikipedia, and 9.6% of what we learn from Wikipedia can be found in the (larger) set of weblogs. As a sign of how distinct these corpora are, if after we've extracted from 50 million weblog sentences, we double the corpus to 100 million weblog sentences, that gives a 68% increase in the number of unique factoids. However, if we instead extract from 50 million Wikipedia sentences, we will have a 115% increase in the number of unique factoids. Rather than indicating that Wikipedia is a richer source, this shows that the knowledge it contains generally hasn't been encountered in the weblogs.

However, there are two reasons to doubt that the knowl-

[6] Our thanks to an anonymous reviewer who asked just this.

edge found in these corpora is this disjoint: (1) There are many differences in diction and spelling that can lead to distinct factoids with nearly identical meanings. (2) Much of the non-overlap data consists of overly specific facts (often about individuals) and factoids seemingly derived from noisy text.

Figure 2 shows how many of the Wikipedia factoids can be found in ever larger chunks of the (raw) weblog output. Counting all of the Wikipedia output, we see that the gains made are quite slow, which is unsurprising given that the raw output includes facts about many named entities that could not be abstracted by the current set of gazetteers and are unlikely to receive much discussion on weblogs. There are, *e.g.*, rather few weblog posts about Lucius Seneca.

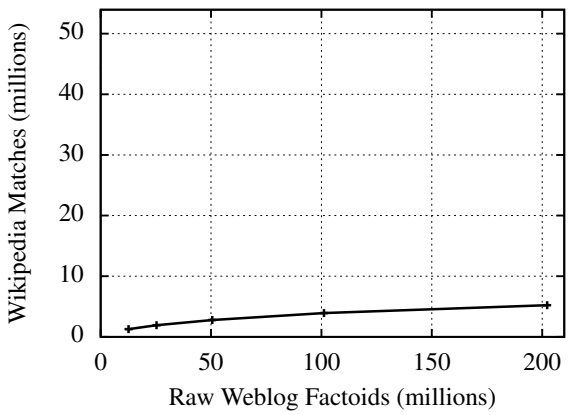

Figure 2: The coverage of Wikipedia factoids by increasingly large amounts of raw weblog output.

Filtering

The real challenge of Web data is to recognize the subset of useful general world knowledge among the chaff. The factoids we wish to discard include those generated from non-English text remaining in the weblogs, those with multiple uncorrected spelling errors, and those mistakenly generated from all sorts of non-text that failed to be preprocessed away.

For this purpose, we introduce a post-processing filter incorporating a parser.[7] We adapted the bottom-up chart parser that EPILOG 2 (Morbini and Schubert 2009) uses for its input to instead parse the logical form output of KNEXT. This is generally a subset of the Episodic Logic used by EPILOG with the addition of colon keywords (such as *:i* for infix well-formed formulas and *:q* for unscoped quantifiers). Note that while we have usually shown the English-like verbalizations of KNEXT's output, what we are parsing and filtering are the logical forms, *e.g.*, *(:i (:q a{n} philosophy.n) encompass.v (:f k (:f plur theory.n)))* – 'A PHILOSOPHY MAY ENCOMPASS THEORIES'.

Parsing KNEXT's output allows us to find when an incorrect parse has led to a syntactically incorrect formula such as *(:i (:q det person.n) (:f Ka break.v))* – 'A PERSON MAY TO BREAK'. This proposition is discarded because the application of a kind-forming operator *(:f Ka ...)* constitutes a term, and the second argument of an infix *(:i ...)* well-formed formula must be a predicate.

The parser also performs lexical checks, requiring that predicates look like potential English words, not line noise. This includes checking that a name contains a part-of-speech suffix (*e.g.*, *person.n* or *sing.v*), is more than one character long, does not contain unlikely punctuation, and either includes a vowel or is in a list of known exceptions such as *CD.n*. Furthermore, to reduce the amount of non-English and misspellings that appears in the output without limiting the use of novel vocabulary, we require that at least 3/4 of the predicates are known words, found in a list consisting of the UNIX dictionary file combined with terms in WordNet 3.0 (Fellbaum 1998) and some manual additions.

[7]While a number of the criteria applied in the filter could be made part of the normal KNEXT extraction process, parsing the logical form is necessarily a postprocessing step, and it was easiest to do all other checks at this point.

Restricting factoids to only using known vocabulary would result in higher quality output but with an unacceptable trade-off in coverage. One advantage of using Web corpora is their currency: neologisms, such as those related to new technologies, are being coined frequently, and a knowledge-extraction system running on the Web will miss much interesting knowledge if it limits itself to recognizable predicates. For instance, we might wish to learn (as KNEXT does) that 'A BLOGOSPHERE MAY EXPLODE WITH DISCUSSION'.

As with less noisy data, errors in the syntactic parsing of English are a common source of bad factoids. For instance, incorrect prepositional phrase attachments in parse trees frequently result in missing arguments, giving incomplete factoids like 'A PERSON MAY FEEL' where what we want to learn is that 'A PERSON MAY FEEL AN EMOTION'. To avoid these incomplete factoids, the filter's parser checks whether a predicate's usage matches the range of arities attested in PropBank (Kingsbury and Palmer 2003) for the corresponding verb. This turns out to be a rather weak restriction given the wide range of possible uses for common verbs, many being uses that KNEXT is unlikely to output. For instance, PropBank includes a use of *say* with no arguments ("Let's assume someone, say John, has been killed"), while KNEXT typically encounters it as a transitive verb. A hand-authored set of corrections to these arity ranges limit such verbs to their common uses.

Since they tend not to convey *general* world knowledge, the filter also removes factoids about named-entities (people, organizations, *etc.*) that could not be abstracted to a more generic kind ('PHILOSOPHER', 'DICTATOR', 'COUNTRY', 'RIVER', *etc.*) using KNEXT's gazetteers. Factoids that contained unclear subjects ('THING', 'THING-REFERRED-TO') are also removed.

As an estimate of the percentage of each corpus that gets removed by these filtering steps, we ran 2000 randomly selected factoids from each corpus through the filter: 567 (28%) of the weblog factoids and 722 (36%) of the Wikipedia factoids were removed. The greater number of factoids thrown out from Wikipedia stems from the greater number of named entities discussed in Wikipedia that could not be

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 3: Instructions for scaled judging.

abstracted and were thus removed by the filter as probably being overly specific.

A small corpus of these filtered KNEXT outputs is being publicly released at http://www.cs.rochester.edu/research/knext/data.

Evaluation

We are interested not only in *what* we can learn from different Web corpora but also the quality of this knowledge: A large but noisy knowledge base will be of little use in reasoning. To measure the quality of knowledge, we must rely on assessments by human judges. We selected 100 propositions uniformly at random from the unfiltered, non-unique[8] output of KNEXT on each corpus. These were shuffled together and their English-like verbalizations were displayed to the judges – in this case, two of the authors – along with the instructions of Van Durme, Qian, and Schubert (2008), seen in Figure 3. Thus the judges did not know which source the factoid they were rating came from nor whether it was among those that would be filtered away.

Some characteristic examples of factoids that were given each rating (agreed on by both judges) are:

1. 'A PERSON MAY HAVE A HEAD'
2. 'A THING CAN BE READABLE'
3. 'A MALE MAY HAVE A CALL'
4. 'CURRENTS CAN BE WITH SOME SURFACE ELECTRODES'
5. 'A % MAY UNDERGO A DEFLATION'

While the highest rated factoid is always true and is at a good level of generality (*person* rather than, say, *male* or *child*), the factoid rated as a 2 is true (some things *are* readable) but is underspecified: What kind of thing is readable? 3 is hard to judge: A person may have a calling or may receive a call, but is the factoid saying either of these? The factoid rated 4 seems a bit too specific (*surface* electrodes) and also a bit vague (*with* them?). The factoid rated 5 we cannot imagine using as knowledge even though we might read a meaning into it: If we take the percent sign to be an adequate stand-in for "percent", we still don't know what it is a percent of. Factoids at each of these ratings can exhibit different problems, but we've found in the past that judges are less likely to agree *what* it is that's wrong with a factoid than *how good* one is (Van Durme and Schubert 2008).

[8]Non-unique output was used to favor more frequently generated propositions. No duplicates were selected.

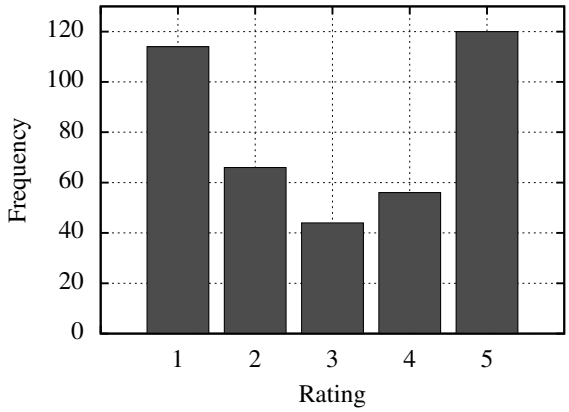

Figure 4: The frequency of ratings assigned to factoids from both corpora.

The distribution of factoid ratings across both corpora can be seen in Figure 4.

The assessments in Table 2 indicate an improvement in the quality of factoids after filtering when compared with the evaluations of the entire, unfiltered set. (For comparison, we estimate that the judgements of KNEXT's output on the Brown corpus, converted to our current rating scale, would have an average rating of around 2.0. This high rating can be ascribed to the accuracy of hand-parses *vs* machine parses.) The evaluations give no indication that the factoids from one Web corpus are of higher general quality than those from the other, with the judges giving roughly the same average rating to each source but Judge 1 slightly favoring those from the weblogs and Judge 2 those from Wikipedia. A larger sample of 300 factoids from each source was evaluated by non-expert judges on Amazon Mechanical Turk. They rated Wikipedia factoids a bit better, and overall assessed quality as higher than the expert judges. For details on this evaluation method, see Gordon, Van Durme, and Schubert (2010).

Approximately equal numbers of factoids from each source passed through the filter: 70 from the weblogs, 71 from Wikipedia. Beyond this filtering, we can also consider only including factoids that are found more than once. Van Durme and Schubert (2008) found that propositions that were extracted at least twice were, on average, judged to better than those extracted only once. However, as extraction frequency continued to increase, the level of judged acceptability did not. We found that for the 200 factoids that were rated, those extracted only once were rated 3.4 on average, while those rated twice or more often were rated 2.79 on average. This is slightly less effective than the other filtering techniques alone. Combining the two, we get a filtered subset of factoids with an average rating of 2.34 *vs* 3.00 overall.

Summary

When extracting general world knowledge, does it matter what machines read? Our findings are that:

1. For both sources, the volume of unique extracted general factoids grows indefinitely, with little sign of leveling off on a logarithmic scale, even after processing of hundreds of millions of weblog sentences.

	Filtered Only			*All*			
	Judge 1	*Judge 2*	*Corr.*	*Judge 1*	*Judge 2*	*Corr.*	*MTurk*
Weblog	2.54	2.52	0.76	3.07	2.98	0.79	2.85
Wikipedia	2.69	2.35	0.71	3.09	2.88	0.76	2.75
Both	2.61	2.43	0.73	3.08	2.93	0.78	2.80

Table 2: Average assessed quality (lower is better – see Fig. 3) for filtered factoids obtained from weblogs and Wikipedia, with Pearson correlation values for the two judges. Last column presents crowdsourced evaluation using Mechanical Turk.

2. Despite the different writing quality in weblogs and Wikipedia, the quality of extracted propositions from those sources are rated about the same by human judges.
3. Use of multiple filtering techniques, such as removal of propositions that fail logical-form parsing, or violate verb arity constraints, or contain many unlexicalized word stems, significantly improves the quality of extracted propositions.
4. Wikipedia-derived general factoids cover only a small fraction of weblog-derived facts and the converse holds also, though the coverage of Wikipedia-derived factoids by weblog-derived factoids appears to grow indefinitely.

Our results suggest that general knowledge extraction from Web-scale text, supplemented with automatic filtering, has the potential to produce large, symbolic knowledge bases of good quality, as judged by people. The next step will be to verify their utility in AI applications.

Acknowledgements

This work was supported by NSF grants IIS-0535105 and IIS-0916599.

References

Banko, M.; Cafarella, M. J.; Soderland, S.; Broadhead, M.; and Etzioni, O. 2007. Open information extraction from the Web. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)*.

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

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

Clark, P., and Harrison, P. 2009. Large-scale extraction and use of knowledge from text. In *Proceedings of the Fifth International Conference on Knowledge Capture (K-CAP 2009)*, 153–60.

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

Gordon, J.; Van Durme, B.; and Schubert, L. K. 2009. Weblogs as a source for extracting general world knowledge. In *Proceedings of the Fifth International Conference on Knowledge Capture (K-CAP 2009)*.

Gordon, J.; Van Durme, B.; and Schubert, L. K. 2010. 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*.

Graff, D.; Kong, J.; Chen, K.; and Maeda, K. 2007. *English Gigaword*. Linguistic Data Consortium.

Kingsbury, P., and Palmer, M. 2003. PropBank: The next level of TreeBank. In *Proceedings of Treebanks and Lexical Theories 2003*.

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

Morbini, F., and Schubert, L. K. 2009. Evaluation of EPILOG: A reasoner for Episodic Logic. In *Proceedings of the Ninth International Symposium on Logical Formalizations of Commonsense Reasoning*, 103–8.

Schubert, L. K., and Tong, M. H. 2003. Extracting and evaluating general world knowledge from the Brown corpus. In *Proceedings of the HLT-NAACL Workshop on Text Meaning*.

Schubert, L. K. 2002. Can we derive general world knowledge from texts? In *Proceedings of the 2nd International Conference on Human Language Technology Research (HLT02)*.

Schubert, L. K. 2009. Language understanding as recognition and transduction of numerous overlaid patterns. In *AAAI Spring Symposium on Learning by Reading and Learning to Read*, 94–6.

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

Van Durme, B.; Michalak, P.; and Schubert, L. K. 2009. Deriving Generalized Knowledge from Corpora using WordNet Abstraction. In *Proceedings of EACL*.

Van Durme, B.; Qian, T.; and Schubert, L. K. 2008. Class-driven attribute extraction. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING-08)*, 921–8.