

Weblogs as a Source for Extracting General World Knowledge

Jonathan Gordon jgordon@cs.rochester.edu

Benjamin Van Durme vandurme@cs.rochester.edu

Lenhart Schubert schubert@cs.rochester.edu

Department of Computer Science
University of Rochester

Supported by NSF grant IIS-0535105.



UNIVERSITY of
ROCHESTER

Open Knowledge Extraction

Enabling human-like understanding and reasoning will require the availability of a great deal of *general knowledge*.

KNEXT (refs: [3, 4]):

- » Abstracts general “factoids” from arbitrary texts – about 2 per sentence.
- » Uses parsing and compositional semantic interpretation rules.
- » Factoids are (underspecified) **logical formulas** (vs tuples in [1])
- » Rendered automatically into approximate English; *e.g.*,
 - » CLOTHES CAN BE WASHED
 - » PEOPLE MAY WISH TO BE RID OF A DICTATOR.

Extracting from Noisy Data

KNEXT has accumulated many millions of factoids, but *human-level intelligent behavior requires many more*, so we turn from traditional corpora to the Web.

Pilot Experiment, using Spinn3r weblog data [2]:

- » Remove/replace obvious non-English text, markup, known abbreviations (*e.g.*, “u r” to “you are”)
- » Process sample: **84 million sentences** (35%) out of 245 million (English) – see Table 1 for extraction statistics.
- » **Fewer factoids per sentence** due to apparent run-on sentences (no punctuation/capitalization) being discarded by the parser.
- » Discard factoids with < 75% known words (using dictionary), *e.g.*,
 - » (ALL MIMSY) CAN BE BOROGOVES

Source	Input Sentences	Raw Factoids	Unique Factoids	Raw per Sentence	Unique per Sentence	Mean Sent. Length
Spinn3r Weblogs	84,301,408	155,405,645	48,785,512	1.84	0.58	16.81
BNC	6,042,908	12,061,685	6,563,622	1.99	1.09	16.28
Web	3,000,736	7,406,371	3,975,197	2.47	1.32	17.05
Brown	51,763	132,113	106,005	2.55	2.05	19.85

Table 1: Factoids found by KNEXT from different sources, before dictionary filtering.

Comparing with Wikipedia

Can text written without the explicit goal of conveying world knowledge offer a similar level of coverage for our knowledge extraction?

NB: We learn **general knowledge** about the world (*men have legs*) rather than **specific information** (*David Bowie was born in 1947*).

Initial Comparison:

- » Identify a random sample of sentential subjects occurring in weblog factoids; *e.g.*, for the factoid **DOORS TO A ROOM MAY BE OPEN -ED**, the subject is **DOORS**.
- » Take the initial (most general) paragraphs about those subjects from Wikipedia, and run through KNEXT, yielding 172 factoids.
- » Check to what extent the Wikipedia-derived factoids are covered by **ever larger sets** of the weblog factoids. See Fig. 1.

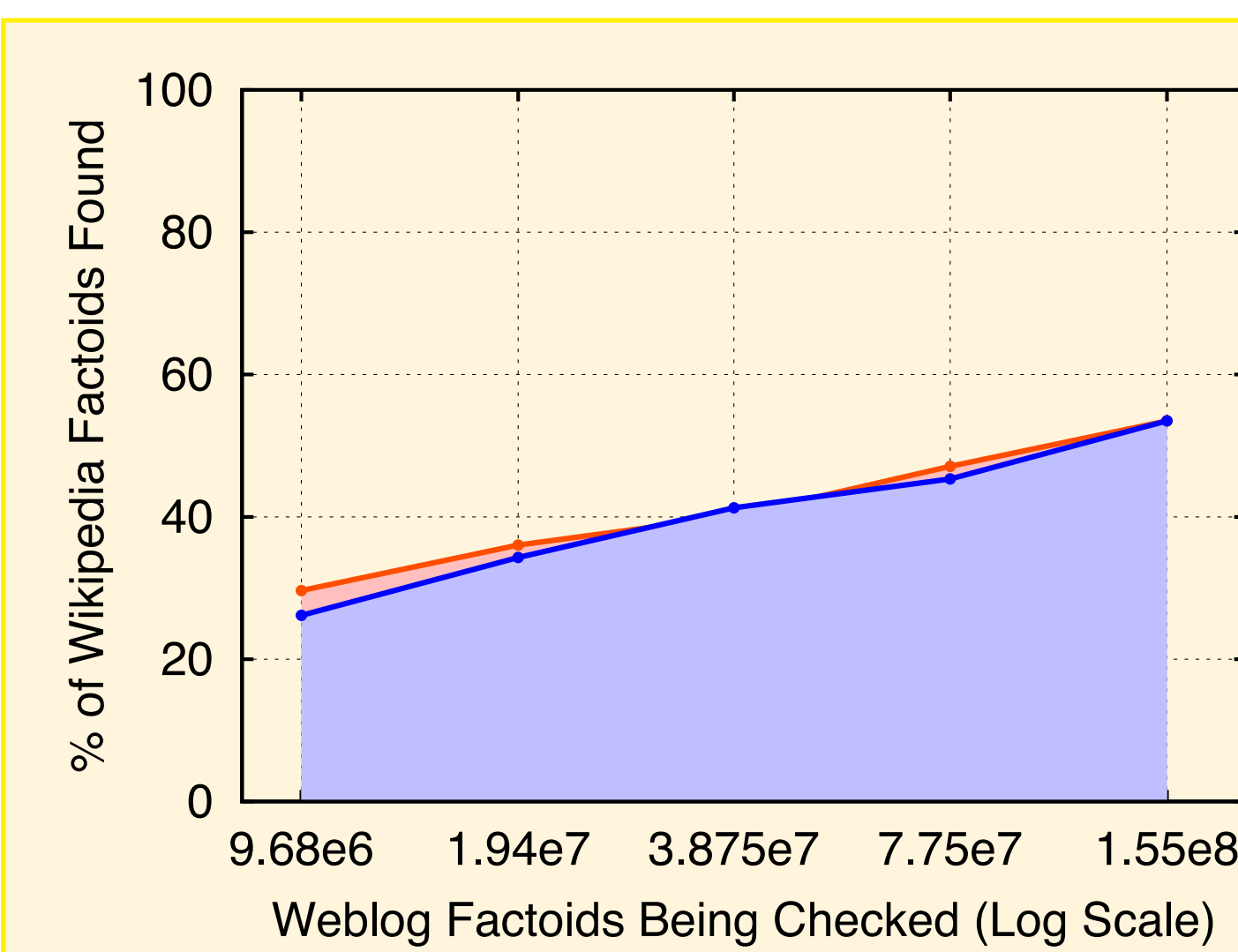


Fig. 1. Coverage of Wikipedia factoids by increasing sets of weblog factoids (for two random shufflings of the weblog factoids).

Some of the Wikipedia factoids not found in the weblog output do occur in possibly equivalent forms, *e.g.*, a factoid containing **A (TIME LINE)** instead of **A TIMELINE**.

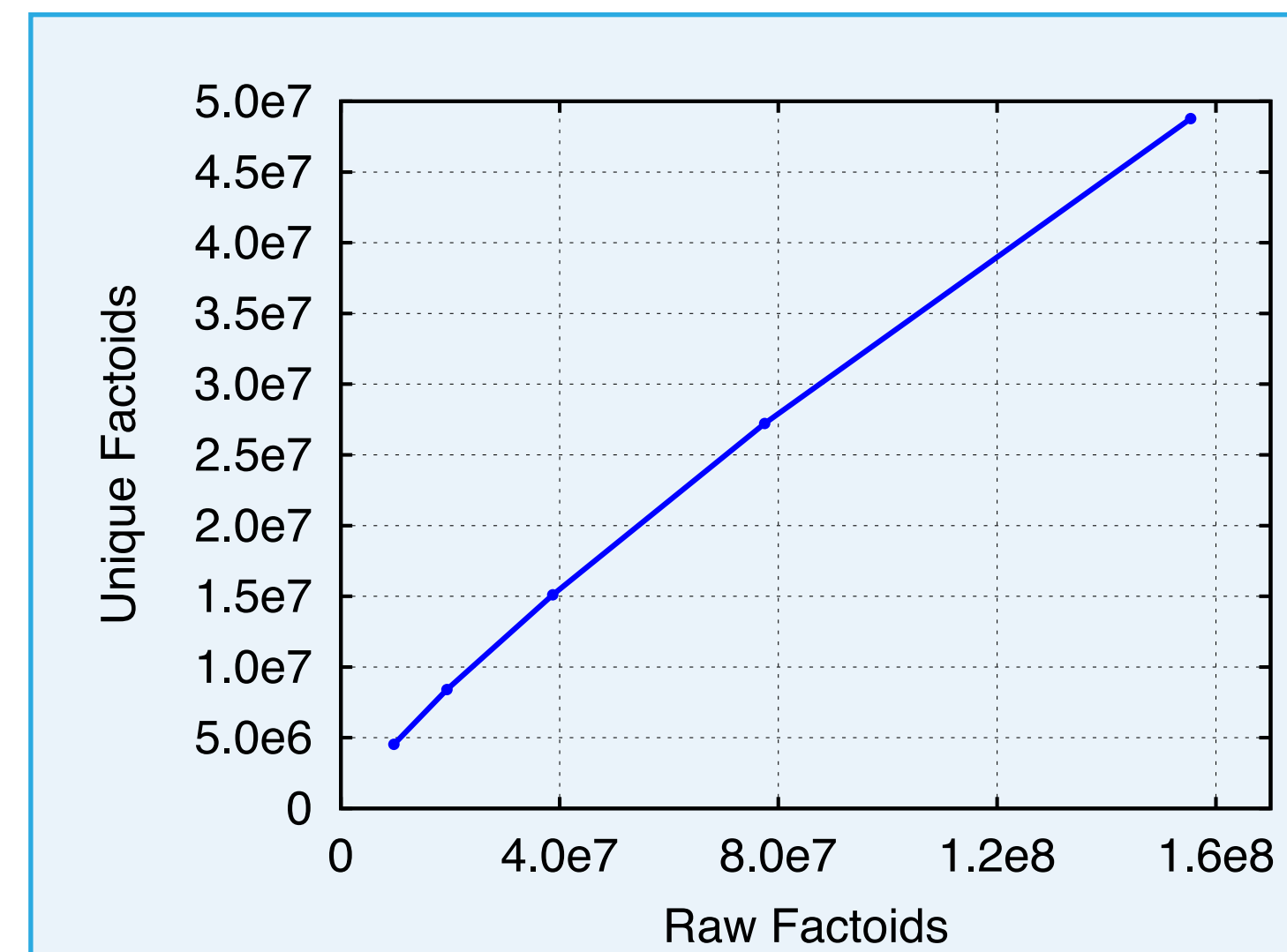


Fig. 2. Rate of growth for unique factoids from the weblog data set as more raw factoids are generated.

There is a basically linear relationship between the production of raw factoids and the number of unique factoids as more text is read from a source like the weblogs.

How many raw factoids would we need to extract from weblogs before we would cover all 172 Wikipedia factoids?



- » Some might never be found
- » Linear extrapolation suggests we would need to produce **18 billion** (raw) factoids from weblog data to reach 100% coverage.
- » This would require 10 billion sentences of weblog text – a very large but possible volume.

Conclusions

- » Lower extraction rates using KNEXT on weblogs suggest casual web text is **harder to parse and learn from**.
- » The majority of factoids derivable from the initial paragraphs of Wikipedia articles can also be obtained from weblogs.
- » Given a great deal of text, weblogs alone *might* be an **adequate source of knowledge** for extraction tools.

Continuing work is on obtaining more complete data on the relative coverage, kinds, and quality of general knowledge obtainable from weblogs vs sources like Wikipedia.

[1] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. *Open Information Extraction from the Web*. In Proc. of IJCAI, 2007.

[2] K. Burton, A. Java, and I. Soboroff. *The ICWSM 2009 Spinn3r dataset*. In Proc. of ICWSM 2009.

[3] L. K. Schubert. *Can we derive general world knowledge from texts?* In Proc. of HLT, 2002.

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

[5] B. Van Durme and L. K. Schubert. *Open Knowledge Extraction through Compositional Language Processing*. In Proc. of STEP, 2008.