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

University of Rochester
Rochester, NY, USA

ABSTRACT

Knowledge extraction (KE) efforts have often used corpora of heavily edited writing and sources written to provide the desired knowledge (*e.g.*, newspapers or textbooks). However, the proliferation of diverse, up-to-date, unedited writing on the Web, especially in weblogs, offers new challenges for KE tools. We describe our efforts to extract general knowledge implicit in this noisy data and examine whether such sources can be an adequate substitute for resources like Wikipedia.

Categories and Subject Descriptors: I.2.6 [Artificial Intelligence]: Learning—*Knowledge Acquisition*

General Terms: Algorithms, Experimentation

1. EXTRACTING FROM NOISY DATA

Enabling human-like understanding and reasoning will require the availability of a great deal of *general* knowledge. We seek to attack this "knowledge acquisition bottleneck" by mining the abundant knowledge that is implicit in phrasal and clausal constructs in texts available in electronic form. To this end, Schubert *et al.* [3, 4] created KNEXT, a system for *open knowledge extraction*. By this we mean one that tries to extract all relations encountered as it "reads" a text, rather than seeking out a small set of targeted relations. Open knowledge extraction is distinct from open *information* extraction (such as [1]) in that it obtains knowledge as symbolic logical formulas rather than tuples of strings. In the case of KNEXT, these logical formulas are automatically translated back into approximate English, giving *factoids* such as "CLOTHES CAN BE WASHED" or "PEOPLE MAY WISH TO BE RID OF A DICTATOR".

KNEXT has accumulated many millions of these factoids, but human-level intelligent behavior requires many more. Thus we turn from traditional corpora to the Web. While a previous experiment [5] used substantial amounts of web text, that work was concerned with comparing KNEXT and TEXTRUNNER [1] – rather than the relative productivity of formal and informal sources – and relied largely on carefully written sources such as Wikipedia and Britannica Online.

As a pilot experiment, we processed the ICWSM 2009 Spinn3r data set [2] of 203 gigabytes of content posted to weblogs.

K-CAP'09, September 1–4, 2009, Redondo Beach, California, USA.
ACM 978-1-60558-658-8/09/09

Source	*Input Sents.*	*Total Raw*	*Total Uniq.*	*Raw (Uniq)/sen.*	*MSL*
Spinn3r	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 [5]	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

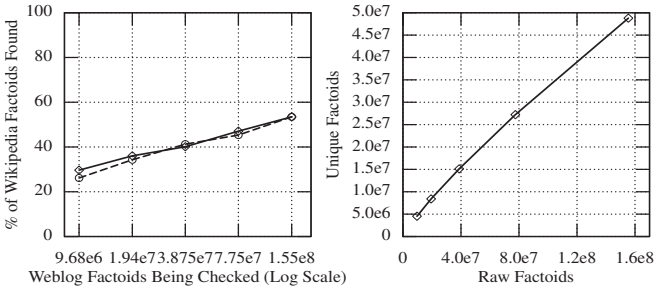
Table 1: Factoids found in different sources. Previously processed corpora were rerun with the current system. MSL is *mean sentence length*.

Preprocessing We removed all HTML tags (marking paragraphs, embedding media, *etc.*), and we elided text found inside tags that indicate non-NL content: ⟨table⟩s, ⟨code⟩, or ⟨pre⟩formatted sections. We also removed large amounts of non-English text, first by removing text that self-identifies as another language (in the content-type of the RSS feeds that were scraped to form [2]). Further filtering was done after extraction – see below. To allow KNEXT to learn from discussion of other web sites, we substituted the text "this website" for all URLs and, likewise, "this email address" for all recognizable email addresses. This is an oversimplification of the ways these are used in casual writing, but it does allow us to conclude from the text "nytimes.com said something cool today" that "A WEBSITE MAY SAY SOMETHING". This is similar in practice to the gazetteer-based abstractions KNEXT uses, *e.g.*, from "René Descartes" to "a philosopher". We also applied a hand-authored set of corrections for common misspellings and the casual mode of online writing, so that, *e.g.*, "u r" is changed to "you are", improving the chances of a correct sentence parse. The preprocessing steps reduced the data set to 26 gigabytes – 245,361,917 recognized sentences.[1]

Extraction We ran KNEXT on a random sample of 84,301,408 sentences, extracting 155,405,645 factoids. However, many factoids are found repeatedly – see Table 1. Observe that the weblog data yield somewhat fewer factoids per sentence than the more formal sources. This cannot be attributed primarily to reduced sentence length, since the differences in mean sentence length are not large. Rather, the lower number of factoids per sentence in the weblogs, and the large number of duplicates[2], reflects the noisy (ungrammatical) nature of much of the writing encountered, including the lack of punctuation and capitalization in many postings, which leads to apparent run-on sentences that are discarded because they exceed the 100-token limit used in our parsing and extraction.

[1]That is, only 12% of the weblog posts are potentially usable English text rather than markup or non-English writing.

[2]The most common duplicates are simple facts about people: "A PERSON MAY THINK", "A PERSON MAY HAVE A LIFE", "A PERSON CAN BE SURE", *etc.*



Figs. 1 and 2: *Left:* Coverage of Wikipedia factoids by increasing sets of weblog factoids (shuffled twice). *Right:* Rate of growth for unique factoids as raw factoids increase.

Filtering In order to remove propositions erroneously generated from remaining non-English text and those generated from sentences with multiple uncorrected spelling errors, we added a post-processing step of checking the factoid verbalizations against a dictionary, discarding those containing less than 75% known words – a cut-off chosen since even non-English sentences may contain English words, but we also want to allow for potentially useful propositions containing neologisms, such as "A BLOGOSPHERE MAY EXPLODE WITH DISCUSSION". An example of a proposition that is rejected by the filter is "(ALL MIMSY) CAN BE BOROGOVES", obtained from a weblog post containing an excerpt of the poem "Jabberwocky" by Lewis Carroll. The filtering described removes 1,192,296 propositions (after the results in Table 1).

2. COMPARISONS TO OTHER DATA SETS

In looking at the output of KNEXT on the weblog data set, we are less interested in measuring the subjective quality of the factoids produced (in the manner of [4, 5]) than in the *types* of knowledge it provides us with, to evaluate the potential usefulness and limitations of extracting from such sources.

As a first comparison, we want to look at Wikipedia, and consider whether text that was written without the explicit goal of conveying world knowledge can offer a similar level of coverage for our knowledge extraction. Since we are interested in general facts about the world (*men have legs*) rather than specific pieces of information (*David Bowie was born in 1947*), Wikipedia may not have a decisive advantage.

To this end, we decided to identify a random sample of sentential subjects occurring in weblog factoids, look up the initial general paragraphs about those subjects in Wikipedia, run these through KNEXT, and then examine the extent to which the Wikipedia-derived factoids were covered by (ever larger portions of) the weblog factoids. More exactly the preparatory steps were this: (1) Select 20 of the weblog factoids, uniformly randomly. (2) Look up the subject of each factoid in Wikipedia, using human judgement about the most reasonable "subject". *E.g.*, for the factoid "DOORS TO A ROOM MAY BE OPEN -ED", the subject is taken to be *doors*, not *rooms*. (3) Use KNEXT to extract propositions from the first 1–2 paragraphs of the article, allowing for articles with short first paragraphs. This resulted in 172 propositions.

To determine the coverage of the sampled Wikipedia knowledge by progressively larger portions of the weblog data, we first checked how many of these Wikipedia factoids could be found (exactly) in 1/16 of the weblog data (shuffled randomly), then in 1/8, 1/4, 1/2, and all of it. The results can be seen in Fig. 1. The complete set of processed weblog data was found to cover 94 (54.6%) of the factoids from Wikipedia. Among those not covered, 25 more (14.5%) seem (by manual inspection) to be present but in slightly different form, *e.g.*, finding "A (TIME LINE)" VS "A TIMELINE", "DISTANCES" VS "A DISTANCE". Of those not found at all, many are not good propositions about the world. One problem with knowledge extracted from Wikipedia is the failure (at the parsing level) to recognize the use–mention distinction. So, a sentence from the "Waste Container" article stating "Common terms are dustbin, rubbish bin, ..." leads to the nonsensical formulation "TERMS CAN BE A BE DUSTBIN". Nonetheless, many of the worthwhile but esoteric propositions from Wikipedia *are* found in the weblog output, *e.g.*, "A CREEL CAN BE A BASKET".

How many raw propositions would we need to extract from weblogs before we would cover all of these? It's quite possible that some Wikipedia propositions would never be found, but simple linear extrapolation of the logarithmic graph suggests that we would need to produce approximately 18 billion (non-unique) propositions from weblog data to reach 100% coverage of those 172 propositions. Given our rate of producing 1.84 raw propositions per sentence, this would mean we would need to process on the order of 10 billion sentences of weblog text – a very large but possibly attainable volume.

Looking at the rate at which the number of unique factoids grows relative to the raw total, we see only a slight fall-off (Fig. 2). Although we find many repeated factoids (only 31% of those generated are unique), there is a basically linear connection between the number of propositions we produce and the number of unique propositions we produce.

3. CONCLUSIONS

The intuition that casually written material on the Web may be less useful for general knowledge mining than more formal sources like Wikipedia is to some extent confirmed by the lower extraction rates we achieved using KNEXT on weblogs. However, our preliminary experiments suggest, somewhat surprisingly, that the majority (more than 50%, perhaps a much higher proportion) of factoids derivable from the initial paragraphs of Wikipedia articles can also be obtained from weblogs. Our continuing work will obtain more complete data on the relative coverage and *kinds* of general knowledge obtainable from weblogs vs sources like Wikipedia.

This work was supported by NSF grant IIS-0535105.

4. REFERENCES

[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.