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.

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 [4])
- 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 BORGHOVES

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.

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.

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, and quality of general knowledge obtainable from weblogs vs sources like Wikipedia.

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

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