Dictionary-based English text compression using word endings

Jeehong Yang and Serap A. Savari
EECS Dept., University of Michigan, Ann Arbor, MI 48109 USA

In this paper, we propose an dictionary-based English text compression algorithm; we name it Star Word Ending (StarWE) because of certain similarities to the StarNT algorithm. Furthermore, StarWE borrows techniques that are used in WRT such as EOL coding, punctuation mark modeling, and n-gram matching. The main difference between StarWE and StarNT is in the division of the external dictionary; StarWE divides it by word endings so that the compressor would be able to obtain some of the tag information.

In our algorithm, all of the words in the dictionary are sorted by frequency, and are grouped using the word ending. Since words in the dictionary are not evenly distributed among the word endings, we split or group the word endings so that each word ending group has a roughly similar number of words. If a word ending group has too many words compared to other groups, we split the group by appending a letter to the word ending. We regroup groups containing a small number of words so that the resulting group contains a reasonable number of words with similar part-of-speech information.

The experimental result is shown in the following table. For almost all the files, StarWE outperforms the state-of-the-art word-based English text compressor WRT by 0.85% when it is combined with PPMZ. However, the performance deteriorates when we combine StarWE with PPMonstr (by 0.45% on average) or by 0.76% on average for PAQ. StarWE requires more time than WRT. This is because StarWE has to calculate the word ending for each word that is not in the frequent word dictionary and we use a brute force searching algorithm. However, our memory requirement was roughly 39% of the memory requirement of WRT.

<table>
<thead>
<tr>
<th>File</th>
<th>WRT+PPMZ</th>
<th>StarWE+PPMZ</th>
<th>WRT+PMM</th>
<th>StarWE+PMM</th>
<th>WRT+PAQ</th>
<th>StarWE+PAQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>alice29</td>
<td>1.80</td>
<td>1.78 (1.26)</td>
<td>1.70</td>
<td>1.71 (-0.30)</td>
<td>1.66</td>
<td>1.68 (-1.01)</td>
</tr>
<tr>
<td>asyoulik</td>
<td>2.05</td>
<td>2.04 (0.07)</td>
<td>1.92</td>
<td>1.95 (-1.91)</td>
<td>1.85</td>
<td>1.90 (-2.32)</td>
</tr>
<tr>
<td>lcet10</td>
<td>1.59</td>
<td>1.59 (-0.03)</td>
<td>1.49</td>
<td>1.49 (-0.23)</td>
<td>1.45</td>
<td>1.45 (-0.42)</td>
</tr>
<tr>
<td>paper1</td>
<td>1.93</td>
<td>1.93 (0.30)</td>
<td>1.76</td>
<td>1.80 (-2.16)</td>
<td>1.68</td>
<td>1.71 (-1.82)</td>
</tr>
<tr>
<td>paper2</td>
<td>1.91</td>
<td>1.87 (2.36)</td>
<td>1.78</td>
<td>1.77 (0.68)</td>
<td>1.73</td>
<td>1.72 (0.51)</td>
</tr>
<tr>
<td>paper3</td>
<td>2.10</td>
<td>2.00 (4.55)</td>
<td>1.93</td>
<td>1.89 (2.01)</td>
<td>1.84</td>
<td>1.83 (0.40)</td>
</tr>
<tr>
<td>paper4</td>
<td>2.31</td>
<td>2.24 (3.01)</td>
<td>2.08</td>
<td>2.10 (-1.05)</td>
<td>1.97</td>
<td>1.98 (-0.48)</td>
</tr>
<tr>
<td>paper5</td>
<td>2.56</td>
<td>2.54 (0.84)</td>
<td>2.30</td>
<td>2.35 (-2.12)</td>
<td>2.20</td>
<td>2.21 (-0.41)</td>
</tr>
<tr>
<td>paper6</td>
<td>2.06</td>
<td>2.02 (1.92)</td>
<td>1.83</td>
<td>1.84 (-0.36)</td>
<td>1.75</td>
<td>1.76 (-0.58)</td>
</tr>
<tr>
<td>Average</td>
<td>1.80</td>
<td>1.78 (0.85)</td>
<td>1.68</td>
<td>1.68 (-0.45)</td>
<td>1.62</td>
<td>1.63 (-0.76)</td>
</tr>
<tr>
<td>Speed</td>
<td>17.00</td>
<td>23.00</td>
<td>4.00</td>
<td>13.00</td>
<td>65.00</td>
<td>90.00</td>
</tr>
</tbody>
</table>