Pipeline

1. Count 5-grams
2. Adjust counts
3. Compute uninterpolated probabilities
4. Interpolate probabilities
5. Compute backoff sums and merge with probabilities
Pipeline

1. Count 5-grams
2. Adjust counts
3. Compute uninterpolated probabilities
4. Interpolate probabilities
5. Compute backoff sums and merge with probabilities
Sorting: Suffix Order

### Suffix Order

<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>Y</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>Y</td>
<td>Z</td>
<td>B</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>Y</td>
<td></td>
</tr>
</tbody>
</table>
# Sorting: Suffix Order

<table>
<thead>
<tr>
<th>Suffix Order</th>
<th>Context Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 4 3 2 1</td>
<td>4 3 2 1 5</td>
</tr>
<tr>
<td>A A A A A</td>
<td>A A A A A A</td>
</tr>
<tr>
<td>C A A A A</td>
<td>A A A A Y</td>
</tr>
<tr>
<td>A Y A B A</td>
<td>C A A A A A</td>
</tr>
<tr>
<td>A Y Z B A</td>
<td>C A B A A Z</td>
</tr>
<tr>
<td>A A A A Y</td>
<td>A Y A B A</td>
</tr>
<tr>
<td>C A B A Z</td>
<td>A Y Z B A</td>
</tr>
</tbody>
</table>

Sorts are performed with TPIE: [https://github.com/thomasmoelhave/tpie](https://github.com/thomasmoelhave/tpie)

Ivan Pouzyrevsky, Mohammed Mediani, Kenneth Heafield
Sorting: Suffix Order

<table>
<thead>
<tr>
<th>Suffix Order</th>
<th>Context Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 4 3 2 1</td>
<td>4 3 2 1 5</td>
</tr>
<tr>
<td>A A A A A A</td>
<td>A A A A A A</td>
</tr>
<tr>
<td>C A A A A A</td>
<td>A A A A A Y</td>
</tr>
<tr>
<td>A Y A B A</td>
<td>C A A A A A</td>
</tr>
<tr>
<td>A Y Z B A</td>
<td>C A B A A Z</td>
</tr>
<tr>
<td>A A A A A Y</td>
<td>A Y A B A</td>
</tr>
<tr>
<td>C A B A Z</td>
<td>A Y Z B A</td>
</tr>
</tbody>
</table>

Sorts are performed with TPIE: https://github.com/thomasmoelhave/tpie
Pipeline

1. Count 5-grams → suffix sort
2. Adjust counts → context sort
3. Compute uninterpolated probabilities → suffix sort
4. Interpolate probabilities → context sort
5. Compute backoff sums and merge with probabilities
Pipeline

1. Count 5-grams → suffix sort
2. Adjust counts → context sort
3. Compute uninterpolated probabilities → suffix sort
4. Interpolate probabilities → context sort
5. Compute backoff sums and merge with probabilities

Avoid two sorts by using memory $\propto$ vocabulary size
Adjusted Counts

\[
c(w_1^n) = \begin{cases} 
\#(w_1^n) & \text{if } n = 5 \text{ or } w_1 = \langle s \rangle \\
|\{v : vw_1^n \in \text{model}\}| & \text{otherwise}
\end{cases}
\]
Discounts

\[
D(c) = \begin{cases} 
0 & \text{if } c = 0 \\
D_1 & \text{if } c = 1 \\
D_2 & \text{if } c = 2 \\
D_{3^+} & \text{if } c \geq 3 
\end{cases}
\]
Interpolating

\[ p_{KN}(w_i \mid w_{i-n+1}^{i-1}) = \]
\[ \frac{c(w_{i-n+1}^i) - D(c(w_{i-n+1}^i))}{\sum_{w_i} c(w_{i-n+1}^i)} + \gamma(w_{i-n+1}^{i-1}) p_{KN}(w_i \mid w_{i-n+2}^{i-1}) \]

where

\[ \gamma(w_{i-n+1}^{i-1}) = \frac{D_1 N_1(w_{i-n+1}^{i-1} \bullet) + D_2 N_2(w_{i-n+1}^{i-1} \bullet) + D_3 N_3(w_{i-n+1}^{i-1} \bullet)}{\sum_{w_i} c(w_{i-n+1}^i)} \]
Backoffs

\[ B(w_1^n) = \frac{1 - \sum_{v: w_1^n \in \text{model}} p(v \mid w_1^n)}{1 - \sum_{v: w_1^n \in \text{model}} p(v \mid w_2^n)} \]
Use MapReduce! [Brants+, 2007]

Limitation of MapReduce: one input stream, one output stream.

$w_1^n$ needs to speak with $w_2^n \implies$ hard to shard.

See BigFatLM for a Hadoop implementation
http://github.com/jhclark/bigfatlm