A: Theory

We consider a unigram as a single word and a bigram as an ordered pair of words in the form $\{w_1, w_2\}$ separated by space. We have two dictionaries Pw and $Pw2$ that contains unigrams and bigrams respectively along with their frequencies. We use the following notations.

Notations	Meaning
Pw[w]	frequency of w in Pw dictionary
Pwseg[w]	frequency of w in Pwseg dictionary (extracted from wseg_simplified_cn)
Pw2[w_1, w_2]	frequency of $\{w_1, w_2\}$ bigram in Pw2 dictionary
V	Total number of different unigrams or bigrams
N	Sum of frequencies
$voc[w]$	Number of different bigrams started by w

There can be different cases that might happen when we analyze any bigram. However, when we analyze the probability of the very first word of the document or when we can not find the first word w_1 of an ordered pair $\{w_1, w_2\}$ in the dictionary Pw, we only use the unigram model to assign the probability of the word.

According to the unigram model, The probability of a word w is given by:

$$P(w) = \frac{Pw[w] + 1}{Pw.N + Pw.V + 1} \tag{1}$$

if it is found in Pw or

$$P(w) = \frac{1}{Pw.N + Pw.V + 1} \tag{2}$$

if it is not found in Pw

However, we use bigram model for every ordered pair $\{w_1, w_2\}$ when w_1 is found in Pw. According to the bigram model, the conditional probability of w_2 given that we already observed w_1 is calculated. There can be any of the following three cases.

(i) Bigram $\{w_1, w_2\}$ in Pw2
In this case there is a connection between two known words. We consider conditional probability for w_2 as:

$$P(w_2|w_1) = \frac{Pw2[w_1, w_2] + 1}{Pw[w_1] + Pw2.voc[w_1] + 1} \tag{3}$$

(ii) Bigram not in Pw2, w_2 in Pw
Although there is no observed connection, both are known words. In this case, we assign the conditional probability based on the frequency of w_2 in Pw.

$$P(w_2|w_1) = \frac{Pw[w_2] + 1}{(Pw[w_1] + Pw2.voc[w_1] + 1)(Pw.N + Pw.V + 1)} \tag{4}$$

(iii) Bigram not in Pw2, w_2 is not in Pw
In this case, w_2 is completely unknown. Instead of assigning a zero probability, we assign probability of w_2 as

$$P(w_2|w_1) = \frac{1}{(Pw[w_1] + Pw2.voc[w_1] + 1)(Pw.N + Pw.V + 1)} \tag{5}$$

B: Data Structure

Data Structure	Description
Pw,Pwseg, Pw2	Dictionaries with frequencies of unigrams (Pw and Pwseg) and bigrams (Pw2)
chart	Dynamic programming Table to store the words
entry	A structure {word, start position, log probability, back pointer}
pq	priority queue of entries, priority= lowest starting

C: Algorithm

The algorithm is given below:

Algorithm 1 Word Segmentation - Part 1

Input: File with non segmented words $Input$,
unigram dictionary Pw,
unigram dictionary $Pwseg$,
bigram dictionary Pw_2

Output: File with segmented words

Initialize $chart = \{\}$ and $maxlen \leftarrow$ longest bigram length
Initialize priority queue with starting position as priority
for l in range(1,...,maxlen) **do**
  $w_1 \leftarrow$ substring input[0,...,l-1]
  if w_1 is number **then**
    calculate $P_{number}(w_1)$ according to the calibrated probability for numbers
    add $\{w_1, 0, P_{number}[w_1], None\}$ into pq
  else if w_1 is found in Pw **then**
    calculate $P(w_1)$ according to unigram model
    add $\{w_1, 0, P[w_1], None\}$ into pq
  else if w_1 is found in Pwseg **then**
    { # We trust on WSEG data 3 times less, so we multiply its probability by 3 }
    calculate $P_{wseg}(w_1) * 3$ according to unigram model
    add $\{w_1, 0, P_{wseg}[w_1] * 3, None\}$ into pq
  else
    { # We then handle unknown words }
    calculate $P_{unknown}(w_1) * length(w_1)$ according to the calibrated probability for unknown words
    add $\{w_1, 0, P_{unknown}[w_1] * length(w_1), None\}$ into pq
if nothing inserted in last for loop **then**
  $w_1 \leftarrow$ input[0]
  calculate $P(w_1)$ according to unigram model
  add $\{w_1, 0, Pw[w_1], None\}$ into pq

while pq not empty **do**
  $item \leftarrow$ highest priority item from pq
  $end \leftarrow item.length + item.start$
  if $item.probability > chart[end].probability$ **then**
    $chart[end] \leftarrow item$
  $nextstart \leftarrow end + 1$
  if w_1 is found in Pw **then**
    for l in range(1,...,maxlen) **do**
      $w_2 \leftarrow$ substring input[nextstart,...,nextstart+l-1]
      calculate $P(w_2|w_1)$ according to bigram model
      add $\{w_2, nextstart, P(w_1) * P(w_2|w_1), item\}$ into pq
    if nothing inserted in last for loop **then**
      $w_1 \leftarrow$ input[0]
      calculate $P(w_1)$ according to unigram model
      add $\{w_2, nextstart, P(w_1) * P(w_2|w_1), item\}$ into pq
  else
    for l in range(1,...,maxlen) **do**
      $w_2 \leftarrow$ substring input[nextstart,...,nextstart+l-1]
      calculate $P(w_2)$ according to unigram model
      add $\{w_2, nextstart, P(w_1) * P(w_2), item\}$ into pq
    if nothing inserted in last for loop **then**
      $w_2 \leftarrow$ input[nextstart]
      calculate $P(w_2)$ according to unigram model
      add $\{w_2, nextstart, P(w_1) * P(w_2), item\}$ into pq
$item \leftarrow$ chart[end]
while $itemisnotNone$ **do**
  add item.word to wordlist
  $item \leftarrow item.backpointer$
print wordlist in reverse order