MAXMATCH

TOKENIZING TEXT WITHOUT WHITESPACE

相撲の世界は大変厳しいものである。

相撲　の　世界　は　大変　厳しい　もの　で　ある　。

MAXMATCH

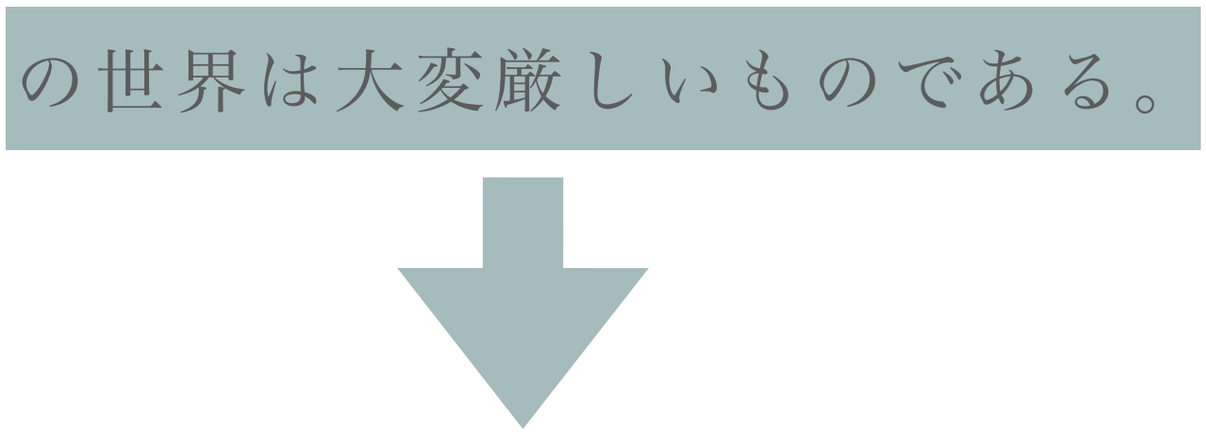
- Using a dictionary, find the longest string of characters that appear in the dictionary.

相撲の世界は大変厳しいものである。

相撲

MAXMATCH

- Using a dictionary, find the longest string of characters that appear in the dictionary.
- Save that string, and repeat on the remainder of the sentence.



相撲　　の

MAXMATCH

- Using a dictionary, find the longest string of characters that appear in the dictionary.
- Save that string, and repeat on the remainder of the sentence.

世界は大変厳しいものである。

相撲　の　世界

MAXMATCH

- Using a dictionary, find the longest string of characters that appear in the dictionary.
- Save that string, and repeat on the remainder of the sentence.

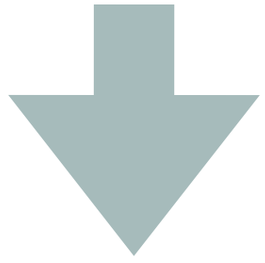
は大変厳しいものである。

相撲　の　世界　は

MAXMATCH

- Using a dictionary, find the longest string of characters that appear in the dictionary.
- Save that string, and repeat on the remainder of the sentence.

大変厳しいものである。

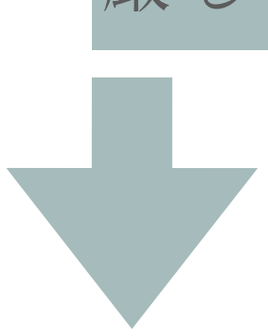


相撲　の　世界　は　大変

MAXMATCH

- Using a dictionary, find the longest string of characters that appear in the dictionary.
- Save that string, and repeat on the remainder of the sentence.

厳しいものである。



相撲　の　世界　は　大変　厳しい

MAXMATCH

- Using a dictionary, find the longest string of characters that appear in the dictionary.
- Save that string, and repeat on the remainder of the sentence.

ものである。

相撲　の　世界　は　大変　厳しい　もの

MAXMATCH

- Using a dictionary, find the longest string of characters that appear in the dictionary.
- Save that string, and repeat on the remainder of the sentence.

相撲　の　世界　は　大変　厳しい　もの　で

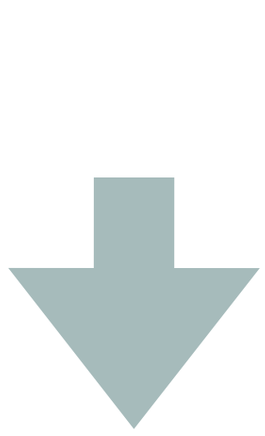
MAXMATCH

- Using a dictionary, find the longest string of characters that appear in the dictionary.
- Save that string, and repeat on the remainder of the sentence.

相撲　の　世界　は　大変　厳しい　もの　で　ある

MAXMATCH

- Using a dictionary, find the longest string of characters that appear in the dictionary.
- Save that string, and repeat on the remainder of the sentence.



相撲　の　世界　は　大変　厳しい　もの　で　ある　。

MAXMATCH

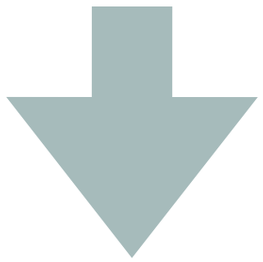
- Using a dictionary, find the longest string of characters that appear in the dictionary.
- Save that string, and repeat on the remainder of the sentence.
- If a word is not encountered in the dictionary, save a single character and repeat on the remainder of the sentence.
- (This algorithm doesn’t handle unknown words well)

相撲　の　世界　は　大変　厳しい　もの　で　ある　。

MAXMATCH

- This approach works decently well for Chinese and Japanese, but not well at all for English.

THETABLEDOWNTHERE

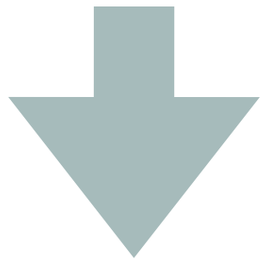


THETA

MAXMATCH

- This approach works decently well for Chinese and Japanese, but not well at all for English.

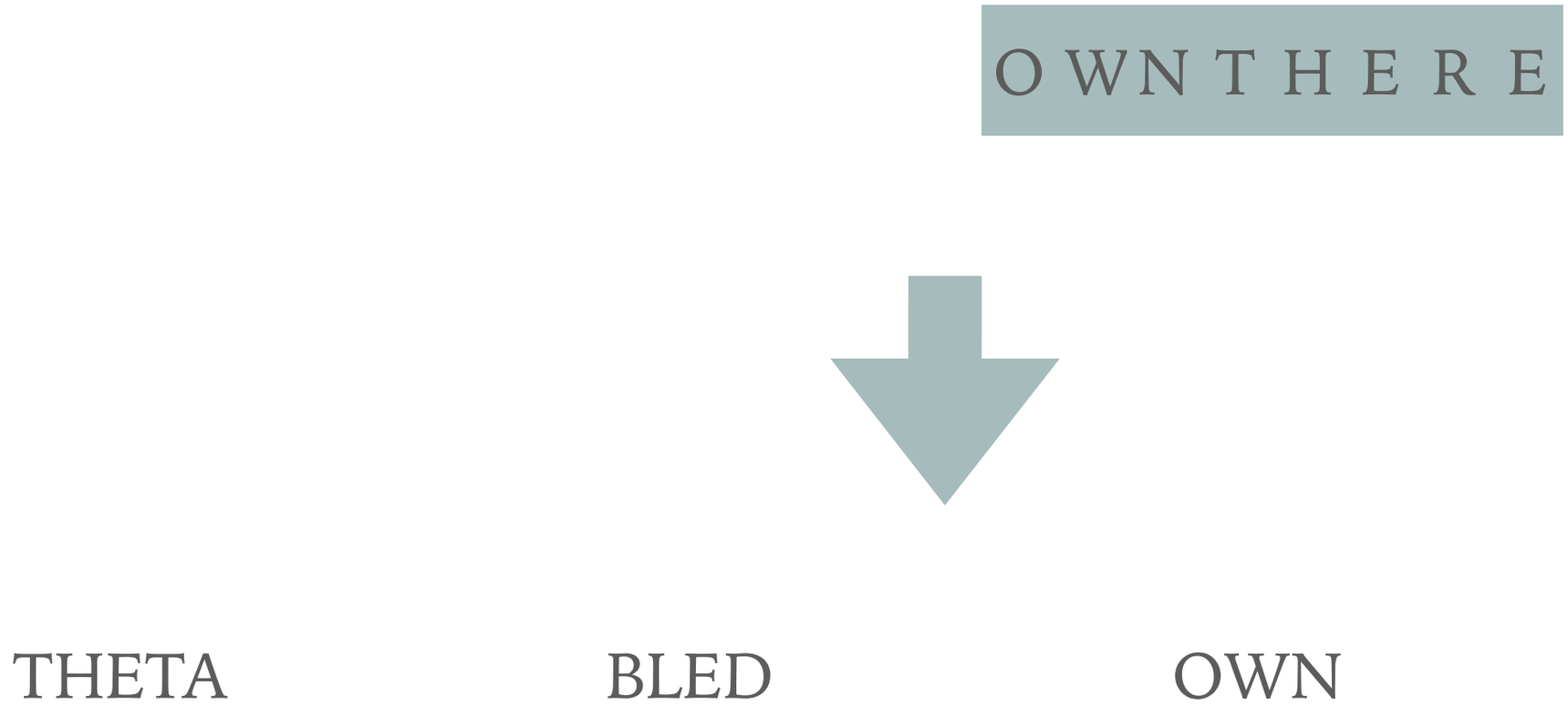
B L E D O W N T H E R E



THETA BLED

MAXMATCH

- This approach works decently well for Chinese and Japanese, but not well at all for English.



MAXMATCH

- This approach works decently well for Chinese and Japanese, but not well at all for English.

THERE

THETA BLED OWN THERE

TOKENIZING TEXT WITHOUT WHITESPACE

- Another approach is to treat this as a classification task.

相	撲	の	世	界	は	大	変	厳	し	い	も	の	で	あ	る	。
B	E	S	B	E	S	B	E	B	I	E	B	E	S	B	E	

- Each character is tagged with one of 4 tags:
  - B - beginning of the word
  - E - end of the word
  - I - inside the word
  - S - standalone word