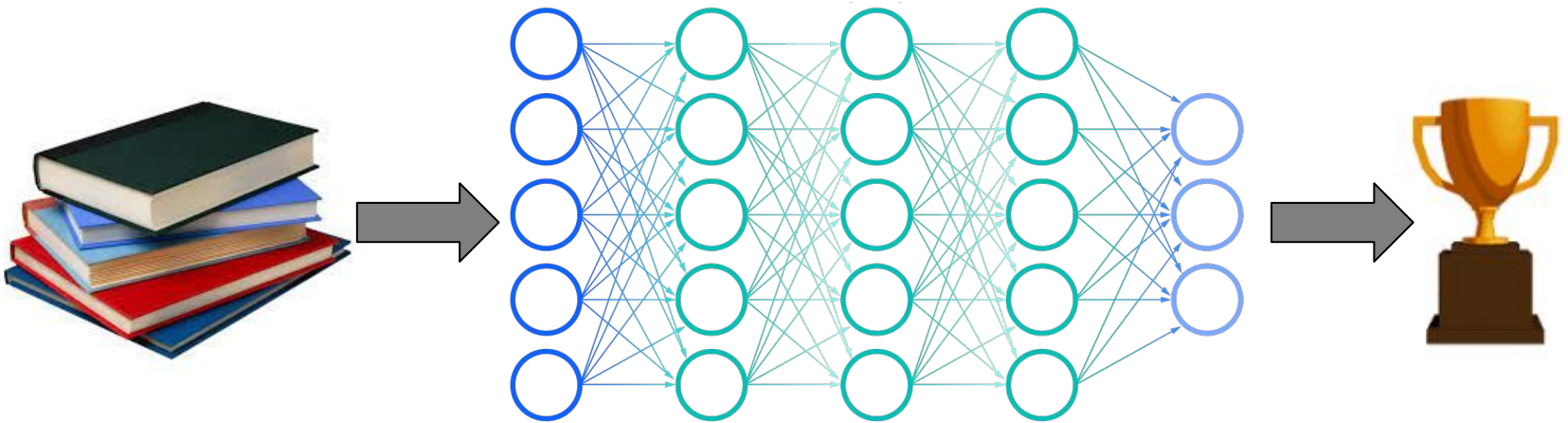
Byte-pair encoding

July 16, 2021

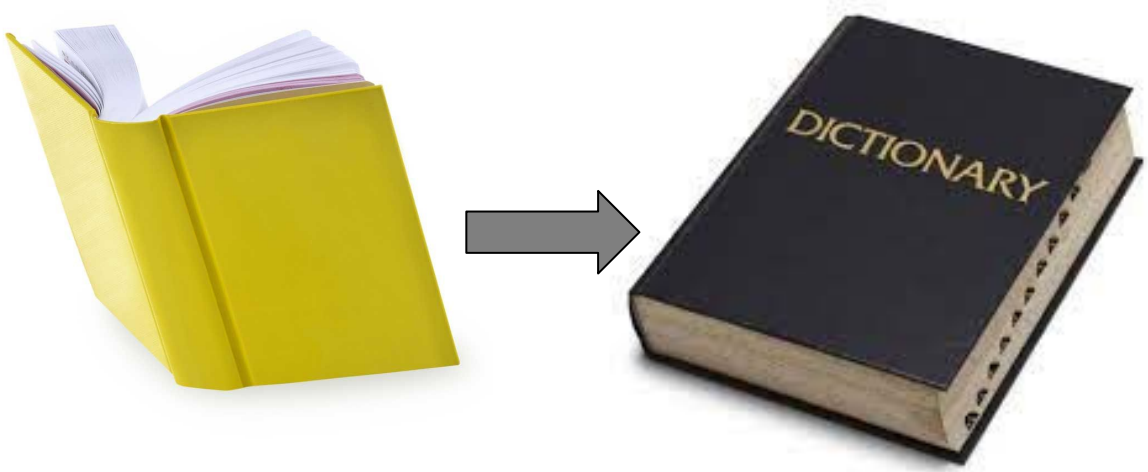
Outline

- How do neural models process text?
- Why is BPE necessary?
- Walkthrough of BPE generation
- Alternatives: SentencePiece & WordPiece, Sequitur, Morfessor

How do neural models read input?

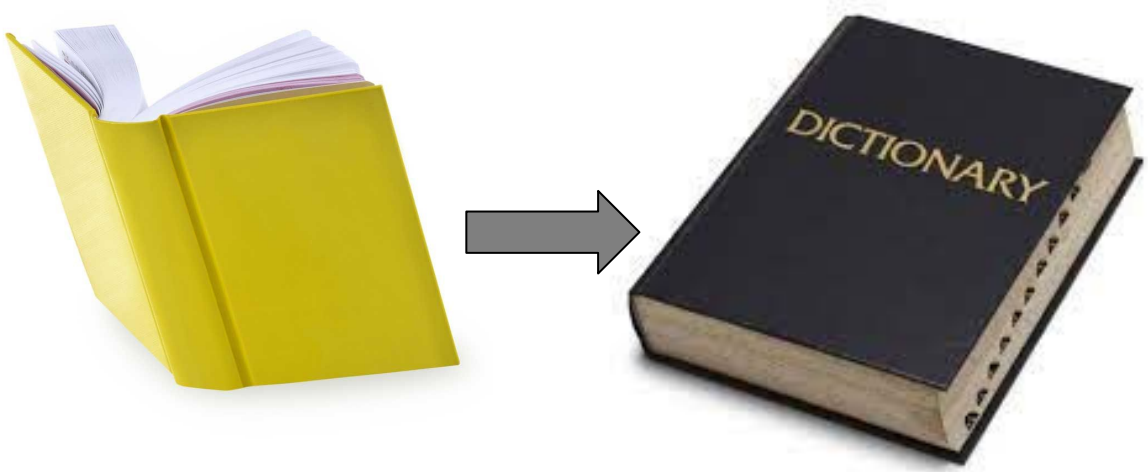


Keep track of all of the words during training



Type	ID
sink	0
representative	1
exclusive	2
shortage	3
confront	4
root	5
technology	6
table	7
forest	8

Keep track of all of the words during training



The birds in the forest heard a sound.
52 1032 19 52 8 27 1941 36 2592

Type	ID
sink	0
representative	1
exclusive	2
shortage	3
confront	4
root	5
technology	6
table	7
forest	8

Abstraction:

The birds in the forest heard a sound.

0	0
1	0
2	0
...	0
51	0
52	1
53	0
...	0
15000	0

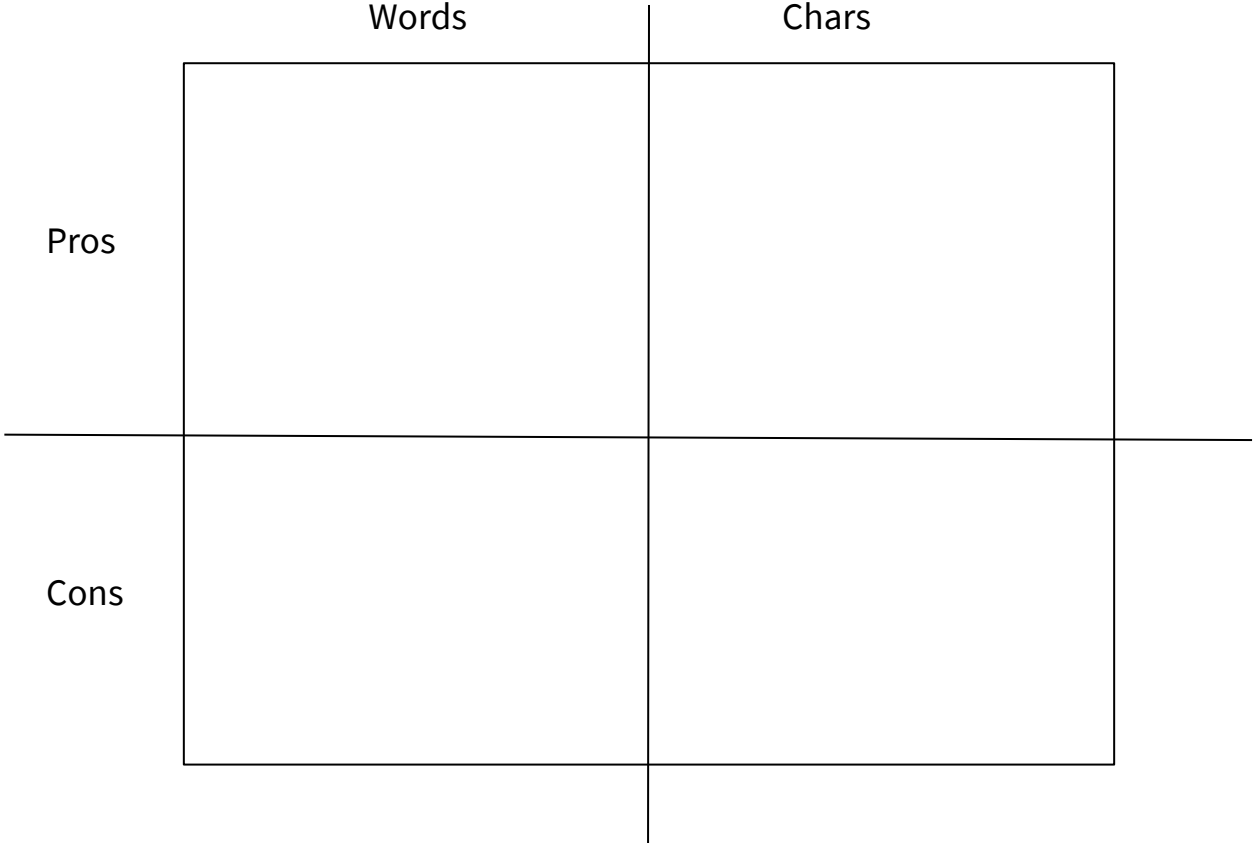
0	0
1	0
2	0
...	0
1031	0
1032	1
1033	0
...	0
15000	0

0	0
1	0
2	0
...	0
18	0
19	1
20	0
...	0
15000	0

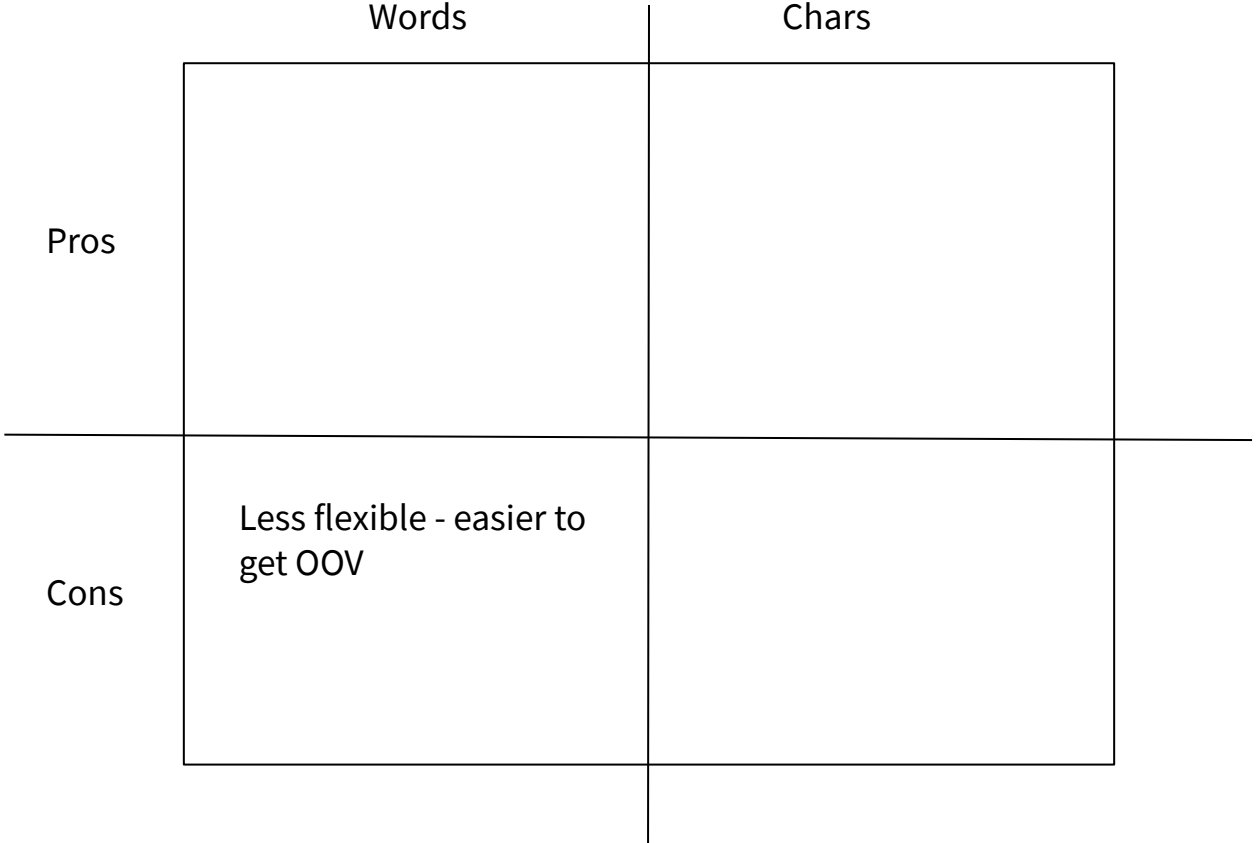
0	0
1	0
2	0
...	0
51	0
52	1
53	0
...	0
15000	0

...

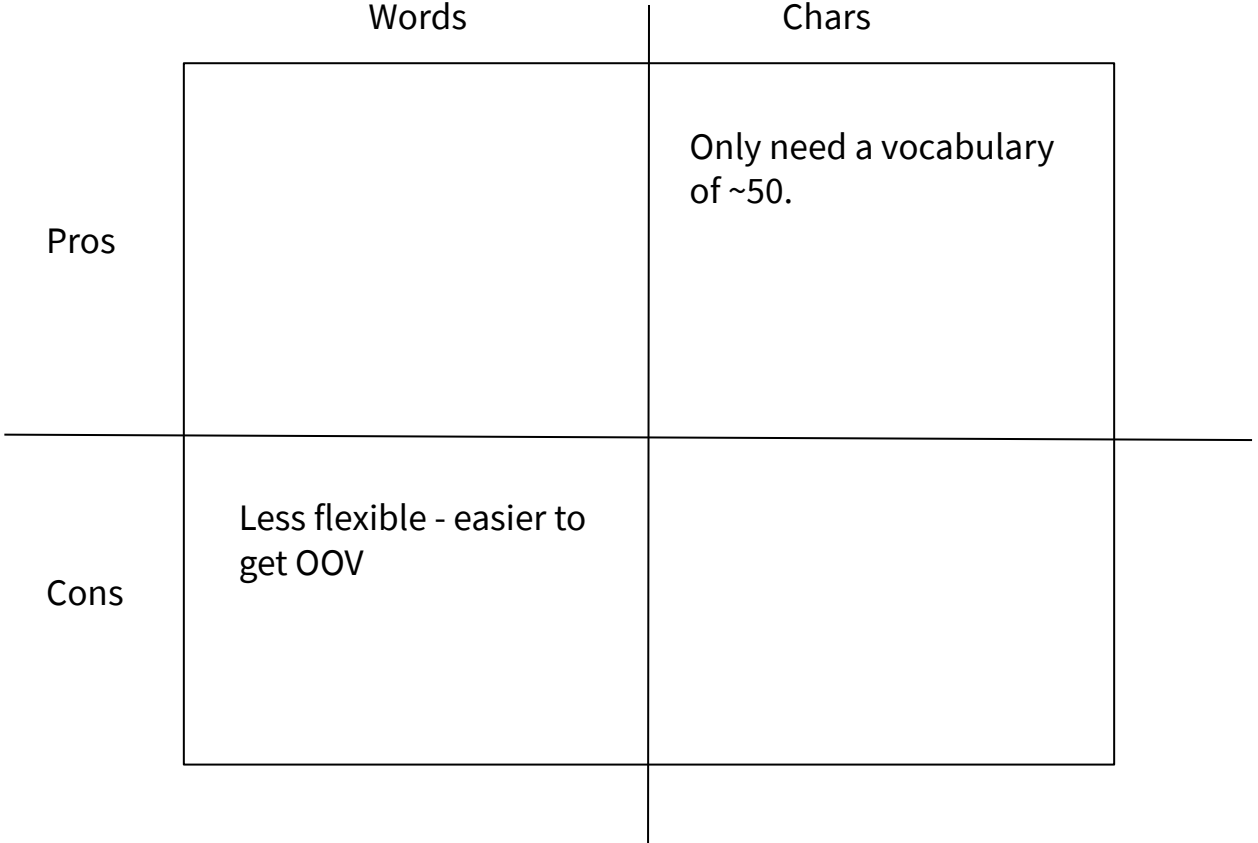
Characters or words?



Characters or words?



Characters or words?



Characters or words?

	Words	Chars
Pros		Only need a vocabulary of ~50.
Cons	Less flexible - easier to get OOV	Letters don’t always represent the same features; context is much larger.

Characters or words?

	Words	Chars
Pros	If seen frequently enough, we can learn very precise embeddings.	Only need a vocabulary of ~50.
Cons	Less flexible - easier to get OOV	Letters don’t always represent the same features; context is much larger.

Subword processing

- Compromise between word-based and character-based models.
- Preserves frequent sequences, but allows some flexibility

The birds in the forest heard a sound: *The bird -s in the forest hear -d a sound*

BPE

- Find common adjacent sequences, build from bottom up.

On the 24th of February, 1815, the look-out at Notre-Dame de la Garde signalled the three-master, the Pharaon from Smyrna, Trieste, and Naples.
As usual, a pilot put off immediately, and rounding the Château d'If, got on board the vessel between Cape Morgiou and Rion island.
Immediately, and according to custom, the ramparts of Fort Saint-Jean were covered with spectators; it is always an event at Marseilles for a ship to come into port, especially when this ship, like the Pharaon, has been built, rigged, and laden at the old Phocee docks, and belongs to an owner of the city.

BPE

- Find common adjacent sequences, build from bottom up.

on the 24th of february , 1815 , the look - out at notre - dame de la garde signalled the three - master , the pharaon from smyrna , trieste , and naples . as usual , a pilot put off immediately , and rounding the château d’ if , got on board the vessel between cape morgiou and rion island . immediately , and according to custom , the ramparts of fort saint - jean were covered with spectators ; it is always an event at marseilles for a ship to come into port , especially when this ship , like the pharaon , has been built , rigged , and laden at the old phocee docks , and belongs to an owner of the city .

BPE

- Find common adjacent sequences, build from bottom up.

on_the_24th_of_february_,_1815_,_the_look_-_out_at_notre_-_dame_de_la_garde_signalled_the_three_-_master_,_the_pharaon_from_smyrna_,_trieste_,_and_naples_.
as_usual_,_a_pilot_put_off_immediately_,_and_rounding_the_château_d'_if_,_got_on_board_the_vessel_between_cape_morgiou_and_rion_island_._immediately_,_and_according_to_custom_,_the_ramparts_of_fort_saint_-_jean_were_covered_with_spectators_;_it_is_always_an_event_at_marseilles_for_a_ship_to_come_into_port_,_especially_when_this_ship_,_like_the_pharaon_,_has_been_built_,_rigged_,_and_laden_at_the_old_phocee_docks_,_and_belongs_to_an_owner_of_the_city_.

e	56
a	45
t	44
o	37
n	31
...	...

BPE

- Find common adjacent sequences, build from bottom up.

o n _ t h e _ 2 4 t h _ o f _ f e b r u a r y _ , _ 1 8 1 5 _ , _ t h e _ l o o k _ - _ o u t _ a t _ n o t r e _ - _ d a m e _ d e _ l a _ g a r d e _ s i g n a l l e d _ t h e _ t h r e e _ - _ m a s t e r _ , _ t h e _ p h a r a o n _ f r o m _ s m y r n a _ , _ t r i e s t e _ , _ a n d _ n a p l e s _ .
a s _ u s u a l _ , _ a _ p i l o t _ p u t _ o f f _ i m m e d i a t e l y _ , _ a n d _ r o u n d i n g _ t h e _ c h â t e a u _ d ’ _ i f _ , _ g o t _ o n _ b o a r d _ t h e _ v e s s e l _ b e t w e e n _ c a p e _ m o r g i o u _ a n d _ r i o n _ i s l a n d _ . _ i m m e d i a t e l y _ , _ a n d _ a c c o r d i n g _ t o _ c u s t o m _ , _ t h e _ r a m p a r t s _ o f _ f o r t _ s a i n t _ - _ j e a n _ w e r e _ c o v e r e d _ w i t h _ s p e c t a t o r s _ ; _ i t _ i s _ a l w a y s _ a n _ e v e n t _ a t _ m a r s e i l l e s _ f o r _ a _ s h i p _ t o _ c o m e _ i n t o _ p o r t _ , _ e s p e c i a l l y _ w h e n _ t h i s _ s h i p _ , _ l i k e _ t h e _ p h a r a o n _ , _ h a s _ b e e n _ b u i l t _ , _ r i g g e d _ , _ a n d _ l a d e n _ a t _ t h e _ o l d _ p h o c e e _ d o c k s _ , _ a n d _ b e l o n g s _ t o _ a n _ o w n e r _ o f _ t h e _ c i t y _ .

e	56
a	45
t	44
o	37
n	31
...	...

th	14
he	11
an	10
nd	8
ar	7
...	...

BPE

- Find common adjacent sequences, build from bottom up.

on_**th**e_24**th**_of_february_,_1815_,_**th**e_look_-_out_at_notre_-_dame_de_la_garde_signalled_**th**e_**th**ree_-_master_,_**th**e_pharaon_from_smyrna_,_trieste_,_and_naples_.
as_usual_,_a_pilot_put_off_immediately_,_and_rounding_**th**e_château_d'_if_,_got_on_board_**th**e_vessel_between_cape_morgiou_and_rion_island_._immediately_,_and_according_to_custom_,_**th**e_ramparts_of_fort_saint_-_jean_were_covered_wi**th**_spectators_;_it_is_always_an_event_at_marseilles_for_a_ship_to_come_into_port_,_especially_when_**th**is_ship_,_like_**th**e_pharaon_,_has_been_built_,_rigged_,_and_laden_at_**th**e_old_phocee_docks_,_and_belongs_to_an_owner_of_**th**e_city_.

e	56
a	45
t	44
o	37
n	31
...	...

th	14
he	11
an	10
nd	8
ar	7
...	...

BPE

- Find common adjacent sequences, build from bottom up.

on_the_24th_of_february_,_1815_,_the_look_-_out_at_notre_-_dame_de_la_garde_signalled_the_three_-_master_,_the_pharaon_from_smyrna_,_trieste_,_and_naples_.
as_usual_,_a_pilot_put_off_immediately_,_and_rounding_the_château_d'_if_,_got_on_board_the_vessel_between_cape_morgiou_and_rion_island_._immediately_,_and_according_to_custom_,_the_ramparts_of_fort_saint_-_jean_were_covered_with_spectators_;_it_is_always_an_event_at_marseilles_for_a_ship_to_come_into_port_,_especially_when_this_ship_,_like_the_pharaon_,_has_been_built_,_rigged_,_and_laden_at_the_old_phocee_docks_,_and_belongs_to_an_owner_of_the_city_.

e	56
a	45
t	44
o	37
n	31
...	...

the	10
an	10
nd	8
ar	7
on	6
...	...

BPE

- Find common adjacent sequences, build from bottom up.

on_**the**_24th_of_february_,_1815_,_**the**_look_-_out_at_notre_-_dame_de_la_garde_signalled_**the**_three_-_master_,_**the**_pharaon_from_smyrna_,_trieste_,_and_naples_.
as_usual_,_a_pilot_put_off_immediately_,_and_rounding_**the**_château_d'_if_,_got_on_board_**the**_vessel_between_cape_morgiou_and_rion_island_._immediately_,_and_according_to_custom_,_**the**_ramparts_of_fort_saint_-_jean_were_covered_with_spectators_;_it_is_always_an_event_at_marseilles_for_a_ship_to_come_into_port_,_especially_when_this_ship_,_like_**the**_pharaon_,_has_been_built_,_rigged_,_and_laden_at_**the**_old_phocee_docks_,_and_belongs_to_an_owner_of_**the**_city_.

e	56
a	45
t	44
o	37
n	31
...	...

the	10
an	10
nd	8
ar	7
on	6
...	...

Some more specifics

- Iterations continue until one of 2 events happens:
  - We reach the pre-specified number of merges (often on the order of 16K, 32K, or 64K).
  - We no longer have any bigrams that occur > 1

Some more specifics

- Iterations continue until one of 2 events happens:
  - We reach the pre-specified number of merges (often on the order of 16K, 32K, or 64K).
  - We no longer have any bigrams that occur > 1
- More merges approaches word-based model; fewer approaches character-based model.

Some more specifics

- Iterations continue until one of 2 events happens:
  - We reach the pre-specified number of merges (often on the order of 16K, 32K, or 64K).
  - We no longer have any bigrams that occur > 1
- More merges approaches word-based model; fewer approaches character-based model.
- Cross-word bigrams are not considered bigrams; likewise, word-internal breaks are indicated by a special symbol: @@ -> This is necessary to re-construct words afterwards.

SentencePiece and WordPiece

- Similar methods to BPE:
- WordPiece has a different selection process - instead of selecting the character bigram with the highest frequency, it chooses the one that maximises the likelihood of the output sequence.
- SentencePiece does everything in Unicode, and works tokenization into the algorithm.

IRRMGP

- Very similar to BPE, SentencePiece, etc.
- Instead of choosing most common bigram, it instead maximizes $|x| * freq(x)$
- Jointly trying to find the longest sequence that is also frequent.

IRRMGP

- Very similar to BPE, SentencePiece, etc.
- Instead of choosing most common bigram, it instead maximizes $|x| * freq(x)$
- Jointly trying to find the longest sequence that is also frequent.

|‘language’| * 50 > |‘es’| * 150
8 * 50 > 2 * 150

language@@ s vs. languag@@ es

MaxLen

- MaxLen finds the longest sequence that has already been seen in the text, and separates it out.

  ‘One language is fun, but many languages are better’

MaxLen

- MaxLen finds the longest sequence that has already been seen in the text, and separates it out.

  ‘One *l**an**guage* is fun, but m**an**y *language* s are better’

Sequitur

Builds a grammar word-by-word:

Sequitur

Builds a grammar word-by-word:

Read in one word - reduce it as much as possible: banana -> baXX; na->X

Sequitur

Builds a grammar word:

Read in one word - reduce it as much as possible: banana -> baXX; na->X

Read in next word, expand grammar.

Morfessor

- Read in all words
- Find segmentation that maximizes the probability of the segments.

Morfessor

- Read in all words
- Find segmentation that maximizes the probability of the segments.

“^languages$” -> “^language + s$”

?

“^languages$” -> “^languag + es$”

Morfessor

- Read in all words
- Find segmentation that maximizes the probability of the segments.

“^languages$” -> “^language + s$”

?

“^languages$” -> “^languag + es$”

Likely “^language + s$” -> removing “-es” would likely create some strange morphemes for other words.