Photo by Raphael Schaller on Unsplash

COMP 3225

Natural Language Processing
2b. Words (Followup)

Prof Leslie Carr lac@soton.ac.uk
University of Southampton

In This Lecture You We Promised …

- "You will learn how to partition a text into tokens effectively using the BPE algorithm"
  - There were a number of questions about this aspect of lecture, with people wanting to understand more about how the BPE algorithm functions in practise
  - I promised to work up a fuller example which is in this slidedeck
    - Working through the BPE tokenization of Peter Pan
    - Comparing to intuitions about naïve word tokenisation

Peter Pan: naïve tokens

Chapter 1 PETER BREAKS THROUGH

All children, except one, grow up. They soon know that they will grow up, and the way Wendy knew was this. One day when she was two years old she was playing in a garden, and she plucked another flower and ran with it to her mother. I suppose she must have looked rather delightful, for Mrs. Darling put her hand to her heart and cried, "Oh, why can't you remain like this for ever!" This was all that passed between them on the subject, but henceforth Wendy knew that she must grow up. You always know after you are two. Two is the beginning of the end.

Of course they lived at 14 [their house number on their street], and until Wendy came her mother was the chief one. She was a lovely lady, with a romantic mind and such a sweet mocking mouth. Her romantic mind was like the tiny boxes, one within the other, that come from the puzzling East, however many you discover there is always one more; and her sweet mocking mouth had one kiss on it that Wendy could never get, though there it was, perfectly conspicuous in the right-hand corner.

The way Mr. Darling won her was this: the many gentlemen who had been boys when she was a girl discovered simultaneously that they loved her, and they all ran to her house to propose to her except Mr. Darling, who took a cab and nipped in first, and so he got her. He got all of her, except the innermost box and the kiss. He never knew about the box, and in time he gave up trying for the kiss. Wendy thought Napoleon could have got it, but I can picture him trying, and then going off in a passion, slamming the door.

Mr. Darling used to boast to Wendy that her mother not only loved him but respected him. He was one of those deep ones who know about stocks and shares. Of course no one really knows, but he quite seemed to know, and he often said stocks were up and shares were down in a way that would have made any woman respect him.

Mrs. Darling was married in white, and at first she kept the books perfectly, almost gleefully, as if it were a game, not so much as a Brussels sprout was missing; but by and by whole cauliflowers dropped out, and instead of them there were pictures of babies without faces. She drew them when she should have been totting up. They were Mrs. Darling's guesses.

Wendy came first, then John, then Michael.

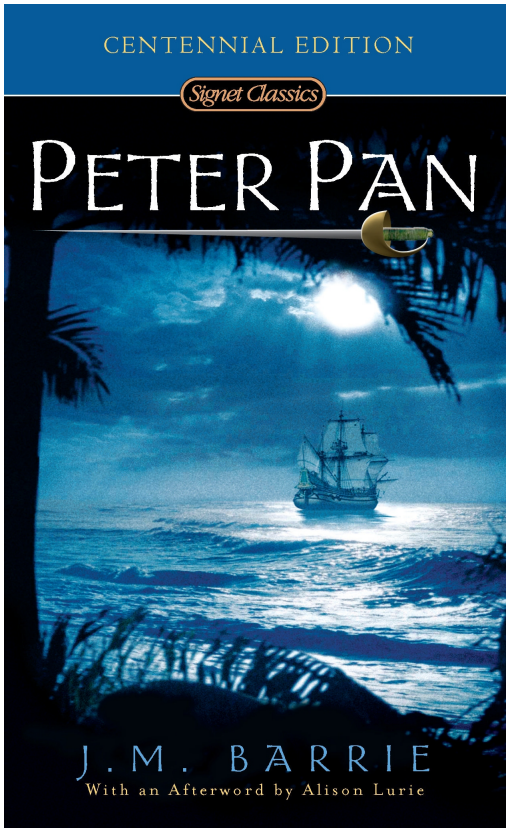
For a week or two after Wendy came it was doubtful whether they would be able to keep her, as she was another mouth to feed. Mr. Darling was frightfully proud of her, but he was very honourable, and he sat on the edge of Mrs. Darling's bed, holding her hand and calculating expenses, while she looked at him imploringly. She wanted to risk it, come what might, but that was not his way; his way was with a pencil and a piece of paper, and if she confused him with suggestions he had to begin at the beginning again.

"Now don't interrupt," he would beg of her.

"I have one pound seventeen here, and two and six at the office; I can cut off my coffee at the office, say ten shillings, making two nine and six, with your eighteen and three makes three nine seven, with five naught naught in my cheque-book makes eight nine seven--who is that moving?--eight nine seven, dot and carry seven--don't speak, my own--and the pound you lent to that man who came to the door--quiet, child--dot and carry child--there, you've done it!--did I say nine nine seven? yes, I said nine nine seven; the question is, can we try it for a year on nine nine seven?"

"Of course we can, George," she cried. But she was prejudiced in Wendy's favour, and he was really the grander character of the two.

"Remember mumps," he warned her almost threateningly, and off he went again. "Mumps one pound, that is what I have put down, but I daresay it will be more like thirty shillings--don't speak--measles one five, German measles half a guinea, makes two fifteen six--don't waggle your finger--whooping-cough, say fifteen shillings"--and so on it went, and it added up differently each time; but at last Wendy just got through, with mumps reduced to twelve six, and the two kinds of measles treated as one.



But this time it's the whole novel....
48077 tokens, 4854 types (50% are unique).

LENGTH	FREQ	TOKEN	LENGTH	FREQ	TOKEN
1	933	a	6	3	action
5	1	aback	6	1	active
9	3	abandoned	6	1	actors
10	1	abandoning	8	1	actually
6	1	abject	5	6	added
6	1	ablaze	8	1	addicted
4	10	able	7	4	address
6	4	aboard	9	2	addressed
5	1	aboil	10	1	addressing
5	108	about	5	1	adept
5	15	above	7	1	adhered
8	2	abruptly	10	6	admiration
7	3	absence	6	2	admire
6	1	absent	7	2	admired
11	4	absolutely	7	1	admires
6	1	absurd	5	2	admit
8	1	absurdly	6	1	admits
6	1	accept	8	5	admitted
8	3	accepted	10	1	admonished
8	1	accident	5	2	adopt
11	4	accompanied	7	2	adopted
12	1	accompanying	6	1	adults
8	1	accorded	7	4	advance
9	2	according	8	4	advanced
7	2	account	9	1	advancing
9	1	accounted	9	3	advantage
8	1	accounts	9	14	adventure
7	1	accurst	10	13	adventures
10	2	accustomed	9	1	advisable
5	1	acorn	6	1	affair
8	1	acquaint	8	2	affected
10	1	acquainted	12	1	affectionate
11	1	acquainting	10	1	affirmative
9	1	acquitted	6	1	afford
6	10	across	8	1	affright …

Naive space-based tokenizing with case folding punctuation deleted, no function-word removal

Peter Pan: BPE tokens

chapter_0 peter_ break s_through _all_ children ,_except_ one,_ grow_up ._they_ soon_ know_that_they_ will_ grow_up ,_and_the_ way_ wendy_ knew_ was_this ._one_ day_ when_she_ was _two_ year s_ old_ she_was_ playing_ in_a_ garden ,_and_she_ pluck ed _another_ flower _and_ ran_ with_it _to_her_mother ._ i_suppose_ she_ must_have_ looked_rather_ delightful ,_for _mrs._darling_ put _her_hand _to_her _heart_and_ cried,_“ oh ,_why_ can't_ you_ remain_ lik e_this_ for_ever !” _this_was _all_that_ passed_ between _them_ on_the_subject ,_but _h enceforth_ wendy_ knew _that_she_ must_ grow_ up._ you _always_ know _after_ you_are_ two . _two_ is_the_ beginn ing_of_the_ end ._of_cours e_they_ lived _at_ 0 4 _[their _house_ number_ on_their _st re et],_and_ until_ wendy_came_ her_ mother_ was_the_ chief_ one ._she_ was_a_ lovely_ lady,_ with_a_ romantic_m ind_and_ such_a_s weet_mocking_mouth ._her_ romantic_m ind_ was_ like_the_ t iny_ box es,_ one_ with in_the_ other ,_that_ come_ from_the_ puzzling_ ea st ,_however_ many_ you_ discover _there_is _always_ one_ more ;_and _her_s weet_mocking_mouth _had_ one_ kiss_ on_it _that_wendy_ could_never_ get ,_though _there_ it_was,_ perfectly_ con sp ic u ous_ in_the_right-hand_ corner ._the_ way_ mr._darling_ won _her_ was_this :_the_ many_ gentle men_ who _had_been_ boys_ when_she_ was_a_ girl_ discovered_ simultaneously _that_they_ lov ed_her ,_and_they_all_ ran _to_her _house_ to_ pro pose_ to _her_ except_ mr._darling ,_who _took _a_ cab _and_ nipp ed_in _first ,_and_so _he_ got _her._he_ got _all_ of _her,_ except_the_ inner most_ box _and_the_ kiss ._he_ never_knew _about_the_ box ,_and_ in _time_he_ gave_ up _try ing_for_the_ kis s._wendy _thought_ na po le on_ could_have_ got_ it,_but_ i_can_ pic tur e_him _trying ,_and_then_ going_off _in_a_ passion ,_s la mm ing_the_ door._mr._darling_ used_to_ boast _to_wendy _that_her_ mother_ not_ only_ lov ed_him_ but_ respect ed_him ._he_was_ one_of_those_ deep _ ones_ who_ know _about_stocks_and_shar es._ of_course_ no_one_ really_ knows ,_but_he_ quite_ seemed_to_ know ,_and_he_ often _said_ stock s_were_ up_and_ shar es_were_ down _in_a_ way_that_ would_have_made_ any_ woman_ respect _him._ mrs._darling_was_ marri ed_in_ whit e,_and _at_first_ she_ kept _the_ book s_ perfect ly, _almost_ gleefully ,_as_if_ it_were_ a_ ga me,_ not_so_much_as_a_ br us sel s_ spr out_ was_ mis sing ;_but_ by_and_by_ whole_ ca u li flowers_ dro pped_out ,_and_instead_ of_them _there_were_ pic ture s_of_ babies_ without_ fac es._she_ drew _them_ when_she_ should_have_been _to tt ing_up ._they_were_ mrs._darling's_ guess es._ wendy_came_ first , _then_ john, _then_ michael ._ for_a_ week_ or _two _after_ wendy_came_ it_was_ doubt ful_ whether _they_ would_be_ able_to_ keep _her ,_as_ she_was _another_ mouth_to_ fe ed._ mr._darling_was_ frightfully_ proud_ of_her ,_but _he_was_ very _hon oura ble ,_and_he_ sat_on_the_ edge_ of_ mrs._darling's_ bed ,_holding_her_hand _and_ calcul ating_ expens es ,_while_she_ looked_at_him_ imploringly _that_was_ not _his_ way ; _his_way_ was_ with_a_ p enc il _and _a_piece_of_paper ,_an _he_had_to_ beg in _at_the_ beginn ing _again._“ now_ don't_ inter rupt ,”_he_would_ be e,_and _two _and_ six _at_the_ offic e ;_ i_can_ cut_off _my_ c of fe e_ at_the_ offic e,_ ,_with_ your_ eight een _and _three_ makes _three_ nine_seven,_ with_f ive_ naught_ nau -- who_is _that_ moving ? -- eight_ nine_seven,_ dot_and_carry_ seven --don't_speak ,_ n who_ came_to_the_ door -- quiet ,_ child-- dot_and_carry_ child-- there,_ you ' ve_ done_ nine_nine_seven ;

35850 tokens, 7868 types (11% are unique).

Iteration continues until no new pairing has freq > 1

Some tokens are whole English words, some are words+punctuation, some are subwords, or words joined with subwords, or multiple words.

Peter Pan: longest BPE tokens

LENGTH	FREQ	TOKEN
46	2	want_always_to_be_a_little_boy_and_to_have_fun
36	1	!”_“look_at_me!”_“look_at_me
35	2	didn't_know_how_she_knew,_she_just_
34	2	would_always_keep_the_window_open_
32	2	feelings_of_the_unhappy_parents_
28	2	chest_of_drawers,_and_peter_
28	2	,”_said_slightly,_“that_
27	2	wendy_and_john_and_michael_
27	2	s_were_out_looking_for_the_
27	2	_a_cowardly_custard.”_“
27	2	,”_replied_the_voice,_“
26	2	would_your_mother_like_you
26	2	slightly_was_the_first_to_
26	2	_they_had_hoped_she_would_
26	2	,”_said_mr._darling._“
26	1	wendy_and_john_and_michael
25	3	medicine_into_nana's_bowl
25	2	hook_or_me_this_time.”_
25	2	go_out_to_dinner_to-night
25	2	fly.”_“i'll_teach_you
25	2	,”_wendy_continued,_“
24	2	you_don't_think_i_would_
24	2	she_would_have_liked_to_
24	2	nest_fell_into_the_water
24	2	ing._at_first_he_thought
24	2	george,”_mrs._darling_
24	2	_his_dagger_at_the_ready
24	2	._there_was_little_sound
24	1	_all_this_time_had_been_
23	2	starkey,”_said_hook,_
23	2	i'll_cast_anchor_in_you
23	2	had_all_his_first_teeth
23	2	colour_of_mother's_eyes
23	2	by_this_time_they_were_
23	2	_the_difference_between
23	2	_sure_i_sometimes_think
23	2	,”_said_slightly,_“
23	2	,”_he_cried,_“i_am_
22	3	:“yo_ho,_yo_ho,_the_
22	3	,”_he_said_a_little_
22	2	what_my_mother_hopes._
22	2	s_were_watching_them._
22	2	mrs._darling_and_nana_
22	2	insters_are_to_be_envi
22	2	go_to_the_office_again
22	2	_mrs._darling_screamed
22	2	_home_under_the_ground
22	2	_about_stocks_and_shar
22	2	._they_all_whipped_off
22	2	._in_the_meantime_the_
22	2	,”_sighed_wendy._“
22	1	sight_of_his_own_blood
22	1	,”_he_whispered,_“
20	5	home_under_the_ground
20	3	,”_cried_wendy,_“
20	2	stick_it_on_with_soap
20	2	mr._and_mrs._darling_
20	2	just_thought_they_did
20	2	building_of_the_house
20	2	bobs,_hammer_and_tong
20	2	_that_he_was_doing_it
20	2	_at_this_moment_that_
20	2	,”_mr._darling_said
20	2	,”_he_said,_“that
20	2	,”_he_said,_“and_
20	2	,”_cried_peter,_“
20	1	_and_john_and_michael
20	3	just_before_the_dawn
20	2	only_a_little_girl._
20	2	on_the_trail_of_the_
20	2	on_the_nursery_floor
20	2	mr._darling,_but_he_
20	2	in_the_morning,_the_
20	2	i_give_you_a_thimble
20	2	hook_or_me_this_time
20	2	home_under_the_trees
20	2	ground_with_an_arrow
20	2	cried_simultaneously
20	2	came_to_the_nursery_
20	2	beginning_of_fairies
20	2	_there_was_a_quiver_
20	2	_tell_the_other_boys
20	2	_home_from_the_party
20	2	_her_children's_mind
20	2	_fear_fell_upon_them
20	2	_at_the_turn_of_the_
20	2	_at_the_foot_of_the_
20	2	.”_“are_you_glad
20	2	._unfortunately_she_
20	2	._the_crocodile_pass
20	2	,”_said_peter._“
20	2	,”_peter_said,_“
20	2	,”_michael_whisper
20	2	,”_cried_the_twins
20	2	,_though_there_were_
20	2	)_describe_mother’s_
...		

Order of byte pairings (1)

	0	40	120	160	200	240	280	320	360	400	440	480
1	e_	it	with	up	se_	cri	't_	ed_to_	s,_	to_the_	bl	quite_
2	_t	ed_	was	were_	ma	on_the_	ould	pir	ite_	bed	what	land
3	_th	sh	she_	ain	qu	itt	_ag	._the_	wor	back	most_	will_
4	_a	es	_they_	peter	of_the_	id	_all	pre	ry	ous	na	yes
5	_h	ir	en_	_that	_as	ound	darl	les	it_was	m_	been	_hand
6	in	gh	ca	aid	ter	in_	dr	con	ant	war	right	ed_in
7	er	et	of_	you_	st_	_they	ay	ght_	with_	like_	ound_	nur
8	t_	om	wend	mo	ever	_there_	su	ent_	pe	knew	_a_s	lad
9	d_	_to_	ll_	ich	ke_	_al	by	ed_the_	lly	came_	other_	gg
10	s_	el	ook	oun	_as_	out_	_ab	other	_this	_him_	leep	_te
10	ou	_ha	_his_	rea	;_	had_	it_was_	childr	if_	mb	mom	urn
12	_the_	_a_	is_	be_	ion	ex	ill	michael	_they_were_	_af	pu	ser
13	,_	ing_	ha	fa	our	ba	ru	,_but	e_h	cried	e._	caus
14	en	oo	_him	il	com	so	fe	a_	don	par	oot	when_
15	y_	you	ck	_have_	sa	me_	e_the_	der	._she_	_at_the_	ing_the_	only_
16	th	for	e_th	es_	dar	ohn	could_	when	_sh	;_and_	_about	;_and
17	._	was_	ur	_her	pp	john	an_	ang	ever_	ie	don't_	_was_
18	on	it_	im	i_	ver	boy	_are_	ga	long	off	gre	becaus
19	ing	ow	ea	to	_at	ind	em	_his	mother	the_	even	who_
20	_an	wi	un	one_	ough	per	_tim	ow_	--	_though	for_the_	_tw
21	wa	bu	he_	_in	do	po	co	mor	self	dd	sc	mi
22	ed	at_	ould_	_she_	_hook	ther_	chil	pirat	sk	_ta	said_	mu
23	o_	is	ear	_her_	_f	__	we	lly_	which	of_cour	fair	call
24	an	be	now	and	we_	ss	icha	_had	l_	nan	fi	_ad
25	or	li	ther	but	peter_	and_	ichael	lea	ch_	by_	ess	befor
26	_s	ere_	di	_their	aid_	look	sel	little_	had	down	over	read
27	ll	le	,_and	wendy	not	so_	al	new	been_	_to_be_	ely	_the_t
28	ar	ai	ro	me	ep	own	ge	pla	_tink	would	pr	ood_
29	_the	_and	_that_	_had_	in_the_	ct	_again	ust_	of_c	,_the_	wer	fl
30	re	ly	ly_	's_	wendy_	for_	_this_	's	your	_he_had_	_se	_time_
31	of	ri	_them	,_wh	way	,_but_	_all_	_to_the_	_think	est_	ul	lov
32	_he_	ve_	us	if	e,_	s._	ey	do_	ld	rough	or_	boys
33	st	_to	ent	lo	_at_	know	what_	_hear	man	come_	_how	pro
34	wh	end	but_	ne	said	no_	mr	fir	car	de	never_	ed,_
35	no	oul	ra	oh	would_	ut	ce	pt	si	_them_	wan	are_
36	er_	pet	le_	ic	ni	_their_	from	gr	night	very_	my_	da
37	at	la	bo	_ar	rom	pa	on_	ig	ust	ought	ure_	_tre
38	_and_	,_and_	_m	out	to_	ong	litt	_then	ke	mrs._darl	sha	it_is_
39	ch	ght	not_	ink	fu	_st	did_	som	s._darl	can	bel	ard
40	_hi	ev	se	go	ce_	._he_	light	ick	op	ely_	ood	never

Order of byte pairings (2)

	520	560	600	640	680	720	760	800	840	880	920	960
0	now_	i_a	go_	in_a_	boys_	bar	e_a	last_	on_his_	went	hear	make_
2	thing_	exc	ion_	_in_	cry	bra	gir	_so_	,_peter	wat	ship	ask
3	ure	,_as_	;_but	i_s	dear	sleep	ble	_from	ster	ence_	lat	girl
4	room	lik	ish	capt	hap	by_the_	near	._it_was_	from_the_	imp	light_	_however
5	way_	;_but_	ing_s	_at_on	ble_	perhap	ree_	remember	have_	her	_think_	ult
6	get	,_who_	des	ter_	n't_	for_a_	listen	michael_	sid	when_she_	ought_	him
7	children	more_	slight	every_	_all_the_	s_and_	ation	_said	spo	is_the_	oh_	cle
8	p_	_thing	vent	ed_up	_tri	y,_	etim	_of_	bla	no,_	do_you	with_m
9	,_and_he_	vo	with_a_	?_	fully	of_them	in_his_	dark	mouth	_there_was_	son	,_she_
10	red	want	moment	cur	ed_his_	int	_as_he_	should_	bro	looking_	_also	fear
10	ed_him	ort	ung	th_	just	could	unt	et_	ed_her_	without_	_to_have_	quest
12	did_not_	_ho	rep	she_was_	well	must_	which_	_happ	sat_	_tootles	nel	perhaps_
13	ge_	mer	_after	one	gu	_ma	um	cha	_he_had	keep	ground	show
14	_then_	_an_	pass	ip	captain	really_	can_	dden	cu	,_and_she_	ed_a_	low
15	ill_	s_were_	ves	mou	flo	_thought	ess_	_a_st	_they_are_	wendy's_	ious	._but
16	won	,_and_the_	brea	air	lago	stea	let	eak	e_to_	crow	good_	you_s
17	ken	w_	ld_	s,_and_	last	e_that_	,_which_	night_	ert	from_	ed_them	fright
18	est	clo	nurser	gave_	_arm	ful_	rock	!_	put_	wait	see_	ner
19	_in_the_	nana	round	,_for	_ask	bet	each_	ile_	ver_	if_you_	]_	ord
20	children_	't	would_have_	ank	_so	er_be	orn	seem	sometim	whis	rang	every
21	ret	under	_away	sme	sten	_hel	saw	_that_they_	it._	ome	i_am_	still_
22	us_	just_	_himself	._there_	even_	ed_her	pirates_	still	mar	_turn	_always_	s_the_
23	say	ome_	med	because_	dist	cc	part	dre	ing_to_	_hat	much	,_for_
24	_through	it_is	with_the_	dea	reme	cl	bea	ream	dog	_hou	john_and_	,_who
25	day	ey_	._they_	_about_	_any_	water	_up	_that_she_	rather_	forgo	put	ry_
26	ed._	of_course_	anc	gi	where_	ous_	_than	codi	land_	face_	del	fel
27	_that_he_	star	_alway	e_that	_he_was_	pped_	ed_it	some	of_a_	eyes	dou	believe_
28	who	was_the_	._darl	:_	fo	ward	_how_	_take_	eer	_time	play	e_they
29	row	br	ine_	str	fle	._then	_answer	of_his_	ance_	rush	_almost_	john_
30	cried_	swer	_that_the_	redsk	fly	of_s	will	_a_little_	before_	exa	ing_up	called_
31	mad	ful	dis	_su	such	repli	know_	hou	diff	ign	e,	_heart
32	first	_toot	beg	_hook_	s_of_	ve	stor	return	crocodi	mis	great_	into_
33	ab	ed_to	i'	mber	min	cla	oft	wo	e_they_	into_the_	bird	_the_s
34	dra	_hea	nib	pan	did	_tell	pirates	lagoon	ss_	_to_him	whisper	._they
35	wind	ies	sp	only	word	fully_	i_have_	,_a	could_not_	belie	nursery_	_these_
36	,_which	ff	was_a_	mrs._darling_	feel	_had_been_	ing,_	follow	give_	wendy_was_	ves_	e_him
37	,_he_	pl	mr._darl	ere	wendy,_	get_	ft	first_	._but_	ance	ppear	_thought_
38	tt	ins	father	jo	went_	ail	some_	_me	window	oh,_	my	made_
39	thing	cro	voic	your_	ight	quick	,_but_he_	cont	ught	our_	_too_	cra
40	let_	llow	bir	,_though	sur	les_	ly,_	_tinker_be	fin	cab	its_	peter's_

Words in both vocabs (N=1162)

a
able
about
act
adventure
after
again
age
aid
air
all
allow
ally
am
among
an
and
answer
any
appear
arm
arrow
art
as
ask
asleep
at
attack
away
awful
ay
babies
baby
back
bad
bag
ball
band
bandages
bar
bark
barred
barrel
base
bask
basket
bath
bathroom
be
beast
beasts
beat
beating
beautiful
beautifully
becoming
bed
been
before
beg
began
begun
behind
being
belay
belief
bell
below
beneath
bent
besides
best
bet
better
between
beyond
big
bird
bit
bitter
black
blade
blanket
blew
blood
blown
bo
board
boast
boat
body
bold
book
born
both
bound
bow
bowl
box
boy
boys
braid
branch
bravely
braves
bread
break
breast
breath
breathing
breaths
brig
bright
brim
broke
broken
brother
brought
brow
brush
brutal
bubble
build
built
bullies
bully
bulwark
bump
burn
burst
busy
but
butter
button
by
c
cab
cabin
call
called
calm
came
can
cannot
cap
captain
care
carefully
careless
carnage
carpet
carried
carry
cast
cat
catch
caught
cautious
cecco
certain
certainly
chain
chair
chamber
character
chatter
cheek
cheer
chief
child
children
chilly
chimney
choose
cigar
cinderella
clang
claw
clear
clever
cleverness
climb
cloak
clock
cloth
cloud
clung
cockiness
cocky
cold
colour
combat
come
comes
comfort
coming
companion
completely
compliment
conceal
conceit
concern
confess
confident
consider
constant
contempt
content
convey
conviction
cook
cookson
corner
correct
cosy
cough
could
count
cover
cradle
craft
crash
crew
cried
crocodile
cross
crow
crowd
cry
curly
curtain
custom
cut
...

common, short words

Naïve words not in BPE list (N=3692)

aback
abandoned
abandoning
abject
ablaze
aboard
aboil
above
abruptly
absence
absent
absolutely
absurd
absurdly
accept
accepted
accident
accompanied
accompanying
accorded
according
account
accounted
accounts
accurst
accustomed
acorn
acquaint
acquainted
acquainting
acquitted
across
action
active
actors
actually
added
addicted
address
addressed

addressing
adept
adhered
admiration
admire
admired
admires
admit
admits
admitted
admonished
adopt
adopted
adults
advance
advanced
advancing
advantage
adventures
advisable
affair
affected
affectionate
affirmative
afford
affright
affrighted
afraid
afterwards
against
aghast
agitated
agitatedly
agitation
ago
agree
agreeable
agreed
agreement
ah

aha
ahoy
aired
airing
alarm
alarmed
alarming
alas
alertness
alf
alight
alighted
alighting
alive
allowed
almost
aloft
alone
along
alongside
aloof
already
alsatian
also
altar
altered
alternately
altogether
always
amazed
amazement
ambition
amends
amiably
amicably
amorous
amount
anchor
angela
angelically

angels
angered
angle
angry
anguish
animal
animals
ankles
announced
announcement
annoy
annoyed
anon
another
answered
answers
anxious
anxiously
anything
apart
apartment
aped
aperture
apiece
apologetic
apologetically
appalled
appalling
appeal
appealed
appealing
appearance
appearances
appeared
appears
applause
approach
arch
arched
ardently

are
aren
argue
argued
arisen
armchair
armed
armful
arms
arose
around
arrange
arrangement
arrangements
arrested
arrived
arrows
artful
artfully
articles
artifice
ascend
ashamed
ashore
aside
asked
asks
ass
assailed
associated
assumed
assure
assured
astir
astonishing
astonishment
astounded
astounding
attached
attacked

attacking
attacks
attempt
attempted
attend
attended
attention
attire
attitude
attracted
attractive
authentic
authority
authors
autograph
automatically
avail
available
avast
avenger
average
averted
avidly
avoid
avoided
await
awaiting
awake
awe
awed
awestruck
awfully
awkward
axes
azores
babes
bacchanalian
backed

backwards
badly
balance
baldness
bales
balls
bandage
bandaged
banging
banks
banns
barbecue
bare
bared
barked
barking
barque
barrels
basement
bashfully
basketful
baskets
bathed
battle
bauble
bay
baying
beached
beaching
beam
beamed
beard
bearded
bearing
bears
beastly
beaten
beauty

beckoning
become
becomes
bedchambers
bedding
bedroom
beds
bedspreads
bedtime
beef
bees
begged
begging
begin
beginning
begins
begirt
begone
begs
behave
behaved
behaving
believe
believed
believing
belle
bellied
bells
belonged
beloved
belt
belted
berries
beseeching
beseechingly
...

long words with common affixes

BPE words not in naïve list (6706)

a_
a_chimney
a_codfish
a_lady
a_m
a_man
a_s
a_twin
ab
able,_
able_
able_to_
abo
about_
ac
acc
accustom
ach
acles
action_
ad
ad_
aded_the_
adventure_
ag
age_
ahawk
ai
aid_
ail
ail_
ain
aint
air,_
air_
aith
ak
al

al_
alian_
alking
all_
ally_
ally_to_
also_
always_
among_the_
an't_
an_
an_un
anc
ance
ance,_
ance._the_
ance._“
ance_
ance_of_
ance_to_the_
ances
anch
and_
and_a_
and_he_
and_his_
and_now
and_the_
and_there_
andon
ang
angu
ank
ank_
answered_
ant
ant_
antly._“

antly_
any_
ap
ar
ar_
ard
ard_
are_
are_the_
are_they
are_you
are_you_
are_you_glad
are_you_ready,_
are_you_s
arm_
ary_
as_
as_a_
as_he_
as_i_can't_be_
as_the_
asked_
ast
at's
at_
at_it
at_on
at_once,_
at_once_
at_the_
at_which
ate_
ated
ated_
ated_by_
ated_them
ately

ately_
ating
ating_
ation
ation_
att
atta
atter
aw
awa
awful_
awfully_
ay,_ay
ay,_ay,_sir
ay_
ays_
b
b_
ba
bab
babies_
baby_
back_
back_and_
back_without
bad_
bad_form
bandag
barbecu
barbecue_
barr
bath-time,_and_
bathroom._“
bb
bber
be?”_“
be_
be_drown

be_our_mother
be_s
be_the_
bea
beach
bear
bears_
beat_the_tom-tom
beaut
beautifu
became_
becaus
because_
because_he_
because_he_is_
because_of_the_
because_the_
because_they_
because_they_had_
become_
bed-
bed._“
bed_
bed_and_
been_
been_s
befor
before_
before_this
before_you_go
began_
began_to_
begg

beginn
beginning_of_fairies
begun_to_
beha
behaved_
behaving_
behind_him
behind_the_
behind_them
being_
being_a_
being_able_to_
bel
believ
believe_
believe_he_
believe_in_fairies
below!”
benea
bent_
bes
beseech
besid
beside_her
besides,_
best_
better_
between_
big_
. . .

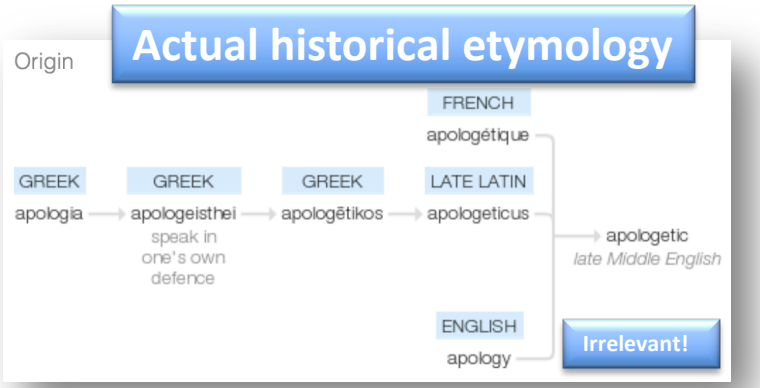
Conjoined word fragments & multiply used morphemes

BPE Sample Tokenisation

- **beseechingly** is tokenized as
  - beseech ingly *two morphologically appropriate subtokens*
- **automatically** is tokenized as
  - _a ut om at ically_. *many morphologically bizarre subtokens*
- **apologetically** is tokenized as
  - ed_apologetically *one-and-a-bit subtokens split across words*
  - this is because it appears twice in the text
    - explained apologetically
    - continued apologetically
  - BPE doesn't know about word boundaries. It just knows that when it sees the string "apologetically" it sees the string "ed_" immediately before.

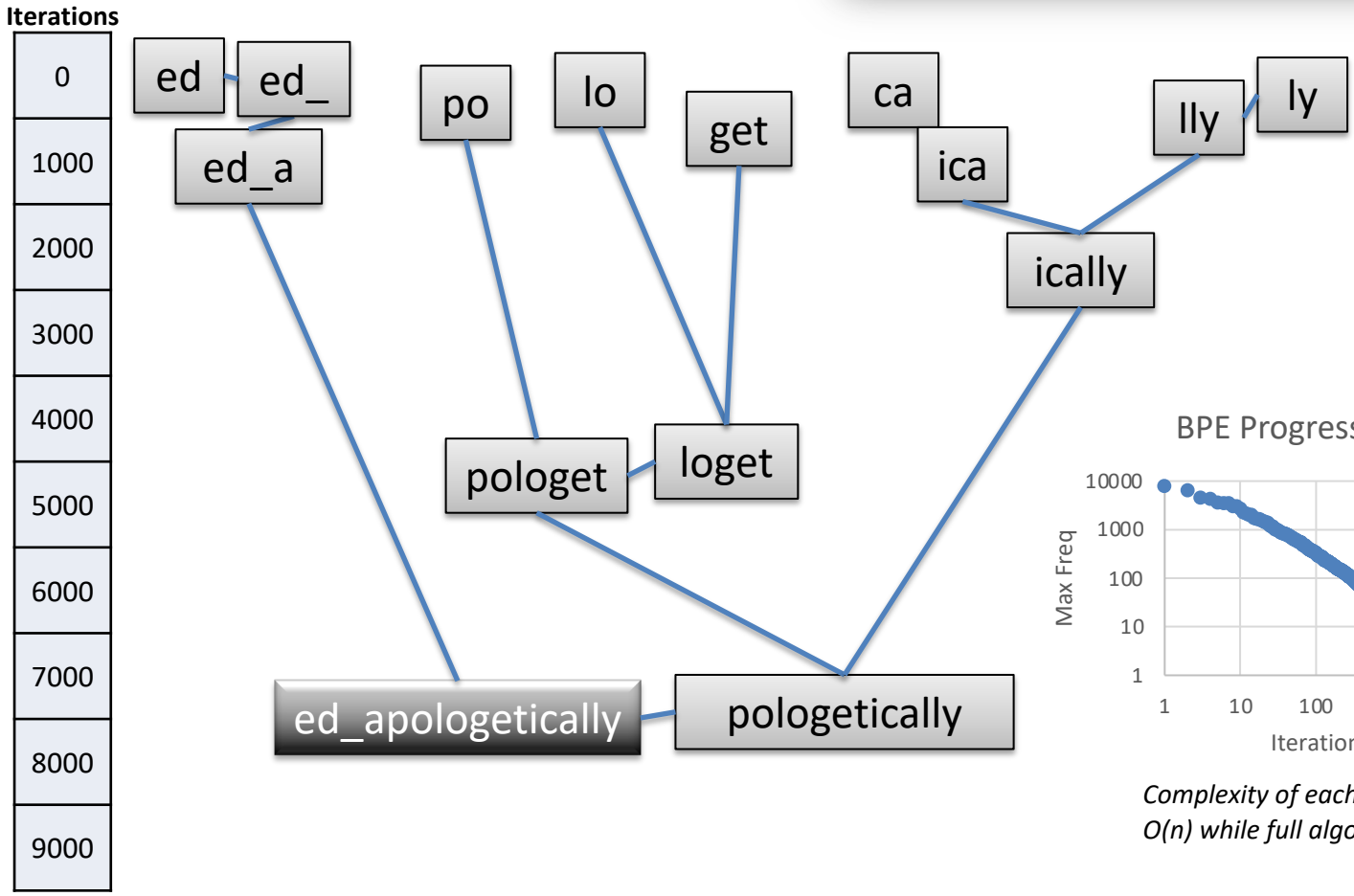
How a token forms

- ed_apologetically



Early BPE iterations deal with short, high frequency tokens

BPE aggregates tokens for statistical, not semantic or morphological, reasons.



Complexity of each iteration is O(n) while full algorithm O(n^2)

SentencePiece

SentencePiece was run on the WikiMatrix (1M+ words, multilingual), and all the English tokens (lowercase alphabetic) were examined and compared with a standard UNIX dictionary (0.25M words). About 6k complete words matched, but of the remainder I examined a dozen to see how well the tokens decompose uncoded words:

abbreviation	abbre + v + iation
apologetic	apolog + etic
behaviorist	behavior + ist
clairvoyant	cl + air + vo + yan + t
beseeching	bes + ee + ching
fragmentation	fragment + ation
glacification	glac + ification
gladiatorial	gl + ad + iat + orial
haberdashery	ha + ber + da + sh + err + y
jackal	jack + al
manipulatively	manip + ulative + ly
pseudopod	pseud + op + od
quadrangle	quadr + angle

A much more useful set of morphemes is generated, although still appears to be less complete than a human stemming.

NB The tokenizing process had been constrained to avoid putting a word break anywhere other than the start of a token.

Peter Pan: adjusted BPE tokens

ch apter_ 1_ peter_ break s _through _ _a ll_ children,_ except_ one,_ grow _up . _the y_ soon_ know _that _the y_ will_ grow _up, _and _th e_ way_ wendy_ knew _was _this ._ one_ day_ when _s he_ was _tw o_ year s_ old_ she_ was_ playing_ in _a _gard en, _an d_ she_ pluck ed _an other_ fl ower _an d_ ran _with _it _to _h er_ mother._ i _sup pose_ she_ must _ha ve_ looked_ rather_ delight ful,_ for _mr s._ darling_ put _her _hand _to _her _heart _an d_ cried,_ “oh,_ why_ can't_ you_ remain _ like_ this_ for_ ever !” _th is_ was _all _th at_ passed_ between _them _on _th e_ subject,_ but _h enceforth_ wendy_ knew _th at_ she_ must_ grow _up ._ you _alway s_ know _af ter_ you _ar e_ two . _tw o_ is _th e_ beginning_ of _th e_ end._ _of _cour se_ they_ lived _a t_ 14 _[their _hou se_ number_ on _their _st re et], _an d_ until_ wendy_ came_ her_ mother_ was _th e_ chief _on e._ she_ was _a _lov ely_ lady,_ with _a _ romantic _mind _an d_ such _a _swe et_ mocking_ mouth. _h er_ romantic _m ind_ was_ like_ the_ t iny_ box es,_ one_ within _th e_ other, _th at_ come_ from _th e_ puzzling_ ea st, _how ever_ many_ you _discover _th ere_ is _alway s_ one_ mor e; _and _her _swe et_ mocking_ mouth _ha d_ one_ kiss_ on_ it _th at_ wendy_ could_ never_ get , _though _th ere_ it_ was,_ perfectly_ con sp ic uous_ in _th e_ right-hand_ corner._ _th e_ way_ mr._ darling_ won _h er_ was _this : _th e_ many_ gent lem en_ who _ha d_ been_ boys_ when _s he_ was _a _girl _discover ed_ simultane ously _that _the y_ loved _her, _and _they _a ll_ ran _to _her _hou se_ to_ pro pose_ to _h er_ except_ mr._ darling,_ who _took _a _c ab _an d_ ni pped_ in _first, _an d_ so _h e_ got _her. _h e_ got _a ll_ of _her ,_ except _th e_ inner most_ box _and _th e_ kiss. _h e_ never_ knew _about _th e_ box, _an d_ in _tim e_ he_ gave_ up _try ing_ for _th e_ kis s._ wendy _though t_ na po le on _could _ha ve_ got_ it,_ but_ i _can _p ic ture_ him _try ing, _and _th en_ going_ off _in _a _passion ,_ sla m ming _th e_ door._ _mr ._ darling_ used _t o_ boast _t o_ wendy _that _h er_ mother_ not_ only_ loved _him _ but_ respect ed _him. _h e_ was_ one_ of _th ose_ deep _on es_ who_ know _ab out_ stocks _an d_ shar es._ of _cour se_ no_ one_ really_ know s,_ but _h e_ quite_ seemed _t o_ know, _and _h e_ often _s aid_ stock s_ were_ up _an d_ shar es_ were_ down _in _a _way _th at_ would _ha ve_ made_ an y_ wom an_ respect _him ._ _mr s._ darling_ was_ mar ried_ in _wh it e, _and _a t_ first_ she_ kept _th e_ book s_ perfect ly, _al most_ gleefully, _a s_ if_ it_ were_ a _gam e,_ not_ so_ much _as _a _br us sel s_ sp r out_ was_ mis sing ;_ but_ by _an d_ by_ whole_ ca u li flowers_ dropped_ out, _an d_ instead_ of _them _th ere_ were_ pic t ures_ of_ babies_ without_ fac es._ she_ drew _them _when _s he_ should _ha ve_ been _tot ting_ up. _the y_ were_ mrs._ darling's_ guesses._ _ wendy_ came_ first , _th en_ john, _then _michael ._ _for _a _week _or _two _af ter_ wendy_ came_ it_ was_ doubtful_ whether_ the y_ would_ be_ able_ to_ keep _her, _a s_ she_ was _an other_ mouth _t o_ fe ed._ mr._ da was_ very _hon ou rab le, _and _h e_ sat_ on _th e_ edge_ of _mr s._ darling's_ bed, _h while_ she_ looked _at _him _im plor ingly._ she_ wanted _t o_ risk _ it,_ come_ what_ way_ was_ with _a _p enc il _and _a _p iece_ of_ paper , _an d_ if _s he_ conf used _h e_ beginning _again ._ _“now _d on't_ inter rupt ,” _h e_ would_ beg _of _her ._ _“i _ha d_ six _at _th e_ offic e;_ i _can _c ut_ off _m y_ c of fe e_ at _th e_ offic e,_ say _ter _your _ eight een _and _th ree_ makes _th ree_ nine_ seven,_ with _f ive_ naught_ naug seven -- who_ is _th at_ mov ing? -- eight_ nine_ seven,_ dot _an d_ carry_ seven --don't_ speak,_ my_ own -- and _th e_ pound_ you_ l ent _to _th at_ man _wh o_ came_ to _th e_ door -- quiet,_ child-- dot _an d_ carry_ child-- there,_ you 've_ done_ it ! -- did_ i _sa y_ nine_ nine_ seven ?_ yes,_ i _s aid_ nine_ nine_ seven ;

62911 tokens, 5706 types (7% are unique).

Iteration continues until no new pairing has freq > 1

Some tokens are whole English words, some are words+punctuation, some are subwords, or words joined with subwords, or multiple words.

Peter Pan: longest revised BPE tokens

LENGTH	FREQ	TOKEN
18	2	_“i--want--you--
15	2	uncomfortably,_
15	2	respectful.”_
15	2	_“besides,”
15	2	_hunting-ground
14	4	unfortunately_
14	2	piccaninnies,_
14	2	panic-stricken
14	2	make-believe,_
14	2	inexperienced_
14	2	extraordinary_
14	1	passionately._
13	4	night-lights,
13	3	children.”_
13	3	_“wendy,”
13	3	_“peter,”
13	2	sufficiently_
13	2	scandalised._
13	2	respectfully_
13	2	night-nursery
13	2	night-lights_
13	2	make-believe_
13	2	made-believe_
13	2	kish-looking_
13	2	indignantly._
13	2	delightfully_
13	2	courteously,_
13	2	beautifully._
13	1	slightly?”_
13	1	drawing-room;
12	6	frightfully_
12	5	immediately_
12	5	_themselves,
12	3	“wendy,”
12	3	pirates.”_
12	3	opportunity_
12	3	extraordinar
12	3	disappeared_
12	2	“peter,”
12	2	undoubtedly_
12	2	suspicions,_
12	2	spectacles._
12	2	shuddering,_
12	2	scornfully._
12	2	satisfaction
12	2	remember,”
12	2	nevertheless
12	2	mantelpiece_
12	2	letters?”_
12	2	ke-believe._
12	2	frightened._
12	2	evening-gown
12	2	doubtfully._
12	2	custard.”_
12	2	_“tink,”
12	2	_“mermaids
12	2	_“lads,”
12	2	_“john,”
12	2	_“hook,”
12	2	_underground
12	2	_codfish!”
12	2	_cheque-book
12	1	ke-believe,_
12	1	innumerable_
12	1	drawing-room
11	8	everything_
11	6	continued,_
11	5	whispered,_
11	5	ke-believe_
11	5	immediately
11	5	captain,”
11	5	_“yes,”
11	4	understand_
11	4	night-light
11	4	especially_
11	4	difference_
11	4	crocodile._
11	4	_slightly's
11	4	_adventures
11	3	understood,
11	3	mother?”_
11	3	mother.”_
11	3	instruction
11	3	instantly,_
11	3	george,”_
11	3	frightfully
11	3	father,”_
11	3	exclaimed._
11	3	everything,
11	3	day-nursery
11	3	children's_
11	3	before.”_
11	3	beautifully
11	3	_themselves
11	3	_mother,”
11	3	_compliment
11	3	_afterwards
11	3	_admiration

BPE Revised Tokenisation

- **beseechingly** is worse
  - bes eech ingly
- **automatically** is the same
  - _a ut om at ically
- **apologetically** is better
  - _apologet ically,_

Unused Lecture Break

- The following slides were prepared for a second break in the catchup lecture but we did not have time to use them

- They give you the opportunity to (a) calculate the naïve word tokenization of some online resources using the Bash tools from the book and (b) try to interpret those vocabularies.

Discussion Questions

- Using the UNIX commands shown on previous slide
  - tr A-Z a-z , tr –cs a-z '\n' , sort , sort -rn , uniq -c , join -v 1
- and the list of function words at
  - https://secure.ecs.soton.ac.uk/notes/comp3225/STOPWORDS.txt

- Try comparing the content of some political speeches
  - chancellor's budget speech in March 2020 at the start of COVID
    - https://www.gov.uk/government/speeches/budget-speech-2020
  - chancellor's budget speech in March 2019 just before Brexit
    - https://www.gov.uk/government/speeches/spring-statement-2019-philip-hammonds-speech

- Can you see any differences relevant to the political context?
  - also try making wordclouds at wordclouds.com
  - alternatively choose other pairs of documents according to your interests

Unused Discussion Questions

- THIS PAGE LEFT DELIBERATELY BLANK WHILE YOU GO AND TRY OUT THE ACTIVITY ON THE PREVIOUS PAGE

- IT'S JUST FOR INTEREST. YOU DON'T HAVE TO DO IT…

Two Budget Speeches

2020 COVID – support

64 today
42 people
39 bn
38 support
37 budget
35 speaker
35 madam
35 deputy
32 year
30 businesses
28 public
27 years
27 funding
26 economy
25 promised
22 tax
22 government
22 fiscal
20 pay
18 investment
18 house
18 country
18 coronavirus
17 fund
16 world
16 work
16 nhs
16 make
16 let
16 invest
15 obr
15 future
15 economic
14 right
14 response
14 relief
14 prosperity
14 need
14 increase
13 jobs
13 impact
12 rhf
12 rates
12 plan
12 national
12 growth
12 cut
12 change
11 taking
11 t
11 spending
11 ll
11 investing
11 high
11 gets
11 building

2019 BREXIT – future of global britain

34 economy
24 year
24 speaker
24 mr
21 future
18 today
18 deal
16 britain
15 spending
15 billion
14 uk
14 political
14 million
14 content
13 review
13 redacted
12 public
12 investment
11 world
11 british
10 years
10 plan
10 house
10 digital
10 budget
9 productivity
9 people
9 need
9 market
9 last
9 announce
8 jobs
8 government
8 generation
8 forecast
8 fiscal
8 delivering
8 deliver
8 additional
7 place
7 pay
7 opportunities
7 national
7 lower
7 housing
7 growth
7 england
7 announced