Natural Language Engineering (2020), **26**, pp. 375–382
doi:10.1017/S1351324920000145

CAMBRIDGE
UNIVERSITY PRESS

EMERGING TRENDS

Emerging trends: Subwords, seriously?

Kenneth Ward Church

Baidu, USA
E-mail: kenneth.ward.church@gmail.com

Abstract

Subwords have become very popular, but the BERT[a] and ERNIE[b] tokenizers often produce surprising results. Byte pair encoding (BPE) trains a dictionary with a simple information theoretic criterion that sidesteps the need for special treatment of unknown words. BPE is more about training (populating a dictionary of word pieces) than inference (parsing an unknown word into word pieces). The parse at inference time can be ambiguous. Which parse should we use? For example, "electroneutral" can be parsed as electron-eu-tral or electro-neutral, and "bidirectional" can be parsed as bid-ire-ction-al and bi-directional. BERT and ERNIE tend to favor the parse with more word pieces. We propose minimizing the number of word pieces. To justify our proposal, a number of criteria will be considered: sound, meaning, etc. The prefix, bi-, has the desired vowel (unlike bid) and the desired meaning (bi is Latin for two, unlike bid, which is Germanic for offer).

Keywords: Subwords; Word pieces; Tokenization; Morphology; Etymology

1. Desiderata

Subwords/word pieces have become quite popular recently, especially for deep nets. They are used in the front end of BERT (Devlin *et al.* 2018) and ERNIE (Sun *et al.* 2019), two very successful deep nets for language applications. BERT provides the following motivation for word pieces:

> "Using wordpieces gives a good balance between the flexibility of single characters and the efficiency of full words for decoding, and also sidesteps the need for special treatment of unknown words." (Devlin *et al.* 2018)

Subwords are based on byte pair encoding (BPE) (Sennrich, Haddow, and Birch 2016), which borrows ideas from information theory to learn a dictionary of word pieces from a training corpus. Word pieces are being used for a variety of applications: speech (Schuster and Nakajima 2012), translation (Wu *et al.* 2016), as well as tasks in the GLUE benchmark (Wang *et al.* 2018),[c] such as: sentiment, paraphrase, and coreference. Many of these papers are massively cited (more than one thousand citations in Google Scholar).

Some examples of the BERT/ERNIE tokenizer are shown in Tables 1, 2, and 3. These tokenizers are intended to be used on text that is like what they were trained on (often wikipedia and newswire), but many of the examples in this paper are selected from something very different to challenge tokenization with lots of out of vocabulary (OOV) words. We collected a small sample of 10k medical abstracts (1.9M words) from PubMed abstracts. More than 30M abstracts are available for download.[d] Medical abstracts are rich in technical terminology (OOVs).

[a] https://github.com/google-research/bert (Bidirectional Encoder Representations from Transformers).
[b] https://github.com/PaddlePaddle/ERNIE (Enhanced Representation through kNowledge IntEgration).
[c] https://gluebenchmark.com/leaderboard/ (General Language Understanding Evaluation).
[d] https://www.nlm.nih.gov/databases/download/pubmed_medline.html

© The Author(s) 2020. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution, and reproduction in any medium, provided the original work is properly cited.

https://doi.org/10.1017/S1351324920000145 Published online by Cambridge University Press

Table 1. Some examples of the BERT/ERNIE tokenizer

Word	PubMed frequency	Word pieces (Sennrich, Haddow, and Birch 2016)	Better alternative
direction	73	direction	
directional	1	directional	
unidirectional	10	un-idi-re-ction-al	uni-directional
bidirectional	1	bid-ire-ction-al	bi-directional
bidimensional	1	bid-ime-ns-ional	bi-dimensional
electroneutral	10	electron-eu-tral	electro-neutral
neurotransmitter	84	ne-uro-tra-ns-mit-ter	neuro-transmitter
potassium	363	potassium	
dipotassium	7	dip-ota-ssi-um	di-potassium
bipotassium	3	bi-pot-ass-ium	bi-potassium
monopotassium	1	mono-pot-ass-ium	mono-potassium
hexapotassium	1	he-xa-pot-ass-ium	hexa-potassium
schizophrenic	54	sc-hi-zo-ph-ren-ic	schizophrenia − a + ic
schizophrenia	43	schizophrenia	
American	20	american	
Telephone	5	telephone	
Telegraph	0	telegraph	
telephony	0	tele-ep-hony	telephone − phone + phony

Many of the analyses in Tables 1, 2, and 3 are surprising. Consider “electron-eu-tral” and “electro-neutral.” BPE is more about training (how to learn a dictionary of word pieces) than inference (how to parse an OOV into a sequence of word pieces). In this case, the parse is ambiguous. How do we choose between “electron-eu-tral” and “electro-neutral”? We suggest minimizing the number of word pieces.

The examples in Tables 1, 2, and 3 raise a number of engineering and linguistic issues. BPE considers letter statistics, but not risk (variance), sound, meaning, etymology, etc. Many of these other factors are considered important for morphological analysis by various communities for various purposes.

1. Engineering considerations
   (a) *Flexibility and coverage*: See quote from Devlin *et al.* (2018) above.
   (b) *Maximize coverage and minimize splits (and especially risky splits)*: Since every split is risky, it is better to use as few word pieces as necessary, especially for frequent words. It should be possible to represent most (frequent) words with one or two word pieces, and almost no words should require more than three word pieces.
   (c) *Avoid risky splits*: Infixes (word pieces in the middle) are more risky than prefixes and suffixes (word pieces at the ends). Short word pieces are more risky than long

https://doi.org/10.1017/S1351324920000145 Published online by Cambridge University Press

Table 2. The analysis of $x + s$ and $x + ed$ reflects frequency. The more frequent form is more likely to be in the dictionary. Regular inflection is relatively safe, but every split is risky, as illustrated by the surprising analysis for "mediates"

PubMed Freq	Word+s	PubMed Freq	Word+ed
656	doses	7	dose-d
571	forms	1	form-d
506	substrates	2	substrate-d
496	membranes	1	membrane-d
319	cultures	132	culture-d
289	complexes	16	complex-ed
176	diseases	12	disease-d
67	induce-s	1499	induced
84	isolate-s	1146	isolated
50	stimulate-s	546	stimulated
26	activate-s	376	activated
33	enhance-s	348	enhanced
8	media-tes	321	mediated
2	character-izes	310	characterized

word pieces.[e] Splits near the middle of words are more risky than splits near the ends. Overlapping splits such as "telephone − phone + phony" are safer than simple concatenation (especially for carefully chosen pairs of affixes like "phone" and "phony").

(d) *Stability*: Similar words should share similar analyses. Small changes should not change the results much.

2. Linguistic considerations

   (a) *Capturing relevant generalizations*: Morphological analyses should make it easy to identify related words: "bidirectional" and "bidimensional" share a common prefix, with similar sound and meaning (and history); "bidirectional" and "unidirectional" share all but the prefix.

   (b) *Sound*: Word pieces should support grapheme to phoneme conversion.

      (i) "bidrectional" and "bidimensional" start with the prefix "bi-" with a long vowel (not "bid-" with a short vowel).

      (ii) "unidirectional" starts with the prefix "uni-" (not "un-"); again, the two prefixes have different vowels.

      (iii) "ction" is unlikely to be a morpheme because English syllables do not start with "ct."

      (iv) Avoid splitting digraphs like "ph" across different word pieces (as in "tele-ephony").

[e] Thirty-five percentage of the PubMed corpus (by token) makes use of a one- or two-letter word piece. These one- and two-letter pieces cover most the possibilities (all 26 one-letter sequences and 421 of 26^2 two-letter sequences).

Table 3. The analysis of hypo-*x* should be similar to hyper-*x*. Hypertension and hypotension, for example, mean high blood pressure and low blood pressure, respectively. Unfortunately, BERT/ERNIE tokenizer splits many of these into too many pieces, making it difficult to see the similarity

hypo-*x* freq	hypo-*x* word pieces	hyper-*x* freq	hyper-*x* word pieces
67	h-yp-ote-ns-ion	216	hyper-tension
40	h-yp-oca-p-nia	60	hyper-cap-nia
20	h-yp-og-ly-ce-mia	12	hyper-gly-ce-mia
18	h-yp-oven-tila-tion	29	hyper-vent-ilation
18	h-yp-oth-er-mic	11	hyper-ther-mic
10	h-yp-oton-ic	40	hyper-tonic
8	h-yp-oth-yr-oid	15	hyper-thy-roid
5	h-yp-ona-tre-mia	1	hyper-nat-rem-ia
3	h-yp-ovo-lem-ia	2	hyper-vo-lem-ia
3	h-yp-oth-yr-oid-ism	9	hyper-thy-roid-ism
3	h-yp-ore-act-ivity	5	hyper-rea-ct-ivity
3	h-yp-op-lastic	6	hyper-pl-astic
3	h-yp-oo-smo-tic	4	hyper-os-mot-ic
3	h-yp-oki-net-ic	4	hyper-kin-etic
3	h-yp-oa-dre-ner-gic	4	hyper-ad-ren-er-gic
1	h-yp-oton-ici-ty	3	hyper-tonic-ity
1	h-yp-ose-ns-iti-zation	1	hyper-sen-sit-ization
1	h-yp-ores-pon-sive-ness	1	hyper-res-pon-sive-ness
1	h-yp-ores-pon-sive	1	hyper-res-pon-sive
1	h-yp-of-un-ction	5	hyper-fu-nction
1	h-yp-oa-ct-ivity	20	hyper-act-ivity
1	h-yp-oa-ctive	9	hyper-active

(c) *Meaning*: "bi" is from the Latin word for two, unlike "bid," which means something else ("offer"), and has a different etymology (Germanic).[f] Similarly, "uni" is from the Latin word for one, unlike the Germanic "un," which means something else ("not" (for adjectives) or "to do in the reverse direction" (for verbs)).[g]

2. Maximize coverage and minimize splits

As suggested above, it should be possible to represent frequent words with one or perhaps two word pieces. Almost no word should require more than three word pieces. Table 1 shows a

[f] https://www.etymonline.com/word/bid
[g] https://www.etymonline.com/search?q=un-

https://doi.org/10.1017/S1351324920000145 Published online by Cambridge University Press

Table 4. Rare words are split more than frequent words, but too many words are split more than necessary

Freq	Number of pieces								
bin	1	2	3	4	5	6	7	8	9+
256+	746	41	27	10	1	0	0	0	0
128	499	69	60	31	11	2	0	0	0
64	717	170	117	56	31	7	1	2	1
32	908	291	231	148	73	22	3	1	1
16	1215	494	409	259	141	45	12	8	2
8	1592	864	775	509	239	73	32	17	15
4	1940	1455	1277	964	427	183	61	26	19
2	2286	2601	2456	1767	850	376	162	82	79
1	3100	5307	5527	4074	2114	979	459	237	229

Table 5. Since every split is risky, it is better to use as few word pieces as necessary, especially for frequent words. Most words are in the dictionary (83% by token), but 5% are split into three or more pieces, and 2% are split into five or more pieces

Word pieces	Coverage
≤ 1	83%
≤ 2	90%
≤ 3	95%
≤ 4	98%
≤ 5	99%

number of examples such as "neurotransmitter" where the BERT/ERNIE tokenizer violates this limit of three word pieces. When this happens, we believe there is almost always a better alternative analysis.

Tables 4 and 5 report coverage by type and by token. Rare words are split more than frequent words, but too many words are split more than necessary. That is, compare the top line in Table 4 for more frequent words to other lines in Table 4 for less frequent words. The top line has relatively more mass in the first few columns, indicating that more frequent words are split into fewer pieces. That said, there are way too many splits. Hardly any words should require more than three pieces, but 30% (by type) and 5% (by token) have more than three word pieces.

3. Risky business

Every split is risky, but some splits are more risky than others. Table 2 shows a number of examples of regular inflection. This is one of the safer splits, but even in this case, "media-tes" is surprising.

In (Coker, Church, and Liberman 1991), we evaluated 11 splitting processes for use in the Bell Labs speech synthesizer. We found that splits near the middle of a word are more risky than splits

https://doi.org/10.1017/S1351324920000145 Published online by Cambridge University Press

Table 6. All splits are risky, but splits in the middle (compounding) are particularly risky

Process	Good or OK
Dictionary Lookup	98%
Stress Neutral	96%
Rhyming	96%
Compounding	86%

toward the end. (Among other things, splits in the middle are more likely to split digraphs such as "ph" as in "tele-ep-hony.")

1. *Compounding*: air-field, anchor-woman, arm-rest, Adul-hussein.
2. *Stress neutral*: abandon-s, abandon-ing, abandon-ment, Abbott-s, Abel-son.
3. *Rhyming analogy*: Plotsky (from Trotsky), Alifano (from Califano).

See Table 6 for the results of an evaluation. We asked a native speaker to listen to about a hundred examples of each case and label each example as:

1. Good: that is how I would have said it,
2. OK: I probably would not say it that way, but I could imagine someone else doing so, and
3. Poor: I know that is wrong.

In addition to accuracy by the 11 splitting processes, we also reported coverage. The splitting processes were designed (as much as possible) to make more use of the more accurate processes and less use of the less accurate processes.

4. Etymology

How can rhyming be risky? Consider the digraph "ch," which usually sounds like the beginning (or end) of my name, "Church," but not in words that come from Italian such as "Pinocchio." So too, if one did not have the Spanish name "Jose" in the dictionary, then one might try to infer that by rhyming with "hose" (and end up with the wrong number of syllables).

In addition to interactions with sound, etymology can also interact with meaning. Consider the the word "digraph." The prefixes "di-" and "bi-" both mean two, but the former is from Greek and the latter is from Latin. Latin also has a prefix "di-," but this prefix means something else ("away from").

This history of English is a long and complicated story that often starts with the Norman Invasion (1066). For a few hundred years after that, the upper classes spoke French and the masses did not. As a result, English now has two words for many things. The Romance (Latin/French) form is often a bit fancier (higher register) than Saxon equivalent. This is particularly clear for food terms, where the upper classes eat beef, pork, and venison (and the serfs raise/hunt cow, pig, and deer).[h] The Norman Invasion introduced many new words for the powers that be in church[i] and state.[j] Many more examples can be found here.[k]

[h] https://www.quickanddirtytips.com/education/grammar/7-french-food-related-words-that-became-english

[i] Religion, theology, clergy, cardinal, dean, pastor, vicar, novice, etc.

[j] Government, administer, crown, state, empire, realm, reign, royal, authority, sovereign, etc.

[k] https://blocs.mesvilaweb.cat/Subirats/?p=58896

https://doi.org/10.1017/S1351324920000145 Published online by Cambridge University Press

Many of the PubMed terms entered the language starting with the scientific enlightenment (at least 500 years after the Norman Invasion),[l] when it was fashionable to coin new terms based on a "revival" of Greek and Latin. The word potassium entered the language relatively recently (1807).[m]

These new words tend to separate Greek and Latin, but not always. My first employer, AT&T underwent a number of reorganizations over the 20 years that I was there. One of them introduced an interesting new word, "trivest,"[n] when AT&T split itself into three parts, soon after "divestment." This is a misanalysis of "divestment" where "di-" is from Latin (meaning "away from")[o] and not the Greek "two." BERT's analyses of these words are surprising: dive-st, tri-ves-t, dive-st-ment, and tri-ves-tment.

AT&T used to be called American Telephone and Telegraph, but they changed their name to AT&T because the telegraph technology (and even the word) does not have much of a future. Interestingly, though, all three words (American, Telephone and Telegraph) are in the BERT lexicon. One might have expected the BERT lexicon to include frequent words with a future, and exclude infrequent words, especially those without a future.

5. Conclusions

Subwords are extremely popular. Many of the papers mentioned here are massively cited. BPE provides a simple information theoretic method for sidestepping OOVs. The method is currently being used for a wide range of applications in speech, translation, GLUE, etc.

That said, it is easy to find surprising analyses such as "electron-eu-tral." If we introduce an additional constraint, minimize the number of word pieces, then we produce the more natural analysis: "electro-neutral."

While the information theoretic BPE criterion sounds attractive to engineers, our field should make room for additional perspectives. Linguists are taught that sound and meaning are better sources of evidence than spelling. This is not an unreasonable position. We should be concerned by the fact that BPE often produces analyses with the wrong meaning and the wrong sound (wrong vowel, splitting digraphs). Such analyses have obvious implications for grapheme to phoneme conversion. For other applications, modern deep nets are so powerful that they can often overcome such issues in preprocessing, but even so, if we can avoid such issues with simple suggestions such as minimizing the number of word pieces, we should do so.

References

Coker C.H., Church K.W. and Liberman M.Y. (1991). Morphology and rhyming: Two powerful alternatives to letter-to-sound rules for speech synthesis. In *The ESCA Workshop on Speech Synthesis,* pp. 83–86.

Devlin J., Chang M.-W., Lee K. and Toutanova K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pp. 4171–4186.

Schuster M. and Nakajima K. (2012). Japanese and Korean voice search. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5149–5152.

Sennrich R., Haddow B. and Birch A. (2016). Neural machine translation of rare words with subword units. In *Association for Computational Linguistics*, pp. 1715–1725.

Sun Y., Wang S., Li Y., Feng S., Tian H., Wu H. and Wang H. (2019). Ernie 2.0: A continual pre-training framework for language understanding. arXiv preprint, arXiv:1907.12412.

[l] https://www.preceden.com/timelines/258474-scientific-revolution-and-the-enlightenment
[m] https://www.etymonline.com/search?q=potassium
[n] https://www.businessinsider.com/att-breakup-1982-directv-bell-system-2018-02
[o] https://en.wikipedia.org/wiki/Disinvestment_from_South_Africa

https://doi.org/10.1017/S1351324920000145 Published online by Cambridge University Press

Wang A., Singh A., Michael J., Hill F., Levy O. and Bowman S.R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. arXiv preprint, arXiv:1804.07461.

Wu Y., Schuster M., Chen Z., Le Q.V., Norouzi M., Macherey W., Krikun M., Cao Y., Gao Q., Macherey K., Klingner J., Shah A., Johnson M., Liu X., Kaiser Ł., Gouws S., Kato Y., Kudo T., Kazawa H., Stevens K., Kurian G., Patil N., Wang W., Young C., Smith J., Riesa J., Rudnick A., Vinyals O., Corrado G., Hughes M. and Dean J. (2016). Googles neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Cite this article: Church KW (2020). Emerging trends: Subwords, seriously? *Natural Language Engineering* **26**, 375–382. https://doi.org/10.1017/S1351324920000145

https://doi.org/10.1017/S1351324920000145 Published online by Cambridge University Press