Word Embedding by Using Word2Vec and BERT

Github link: https://github.com/Allen-ZKW/NLP_HZAU/tree/task7

Abstract

In this task, we use two algorithms to complete Word Embedding. This two algorithms all try to transform words to vectors, these vectors will reflect the closeness and similarity between two different words

Principle

Both of these two algorithms' target is translating word to type of information which can be understand by computers and programs. We can use 'translate data' to calculate distance between two different words. BERT algorithm takes word in different context into concern, so the result of BERT will include more information. This model is trained by forwards information and afterwards information.

Measure

Data Preparation

```
unzip data/litcovid-trainingdata.zip
```

Environmental Preparation

```
conda create -n NLP_task7 python=3.8
conda activate NLP_task7
conda install pytorch torchvision torchaudio cpuonly -c pytorch
conda install --yes --file requirements.txt
```

Word2Vec

```
python3 Skip_Gram_basic.py
```

BERT

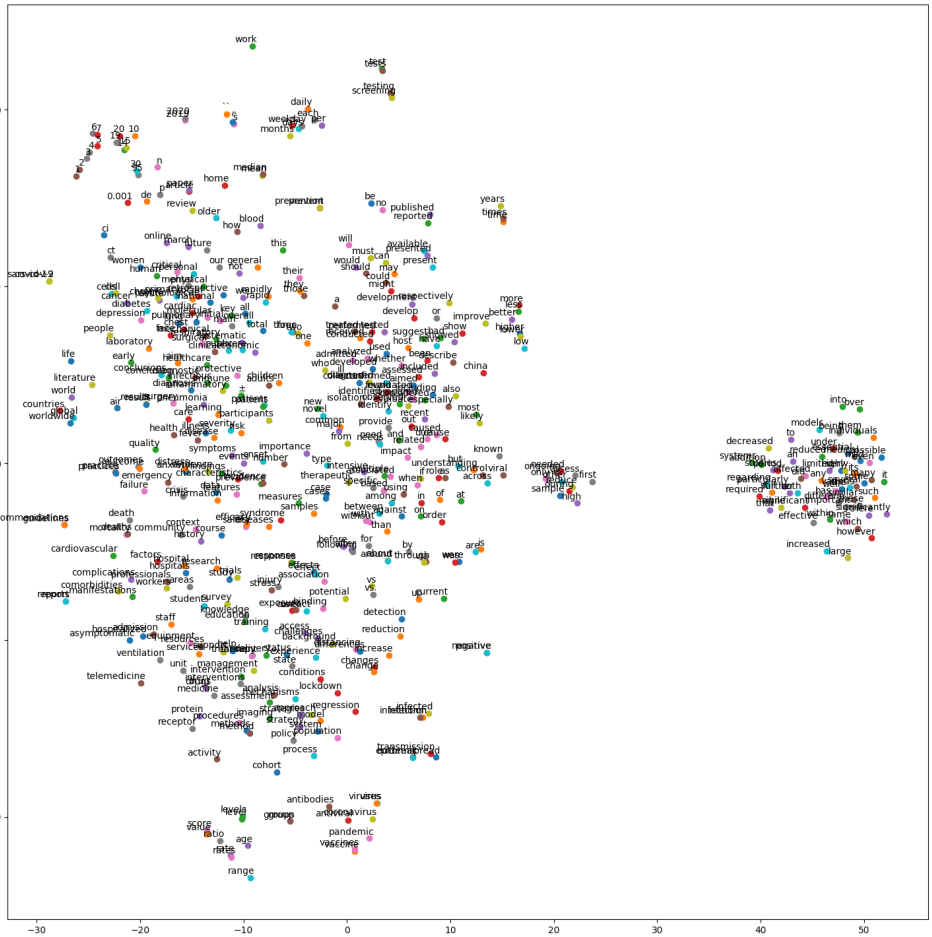
```
python3 Bert_4_Litcovid_WordEmbedding.py
```

Result

BERT Vectors

the 0.356284648418000793 -0.5250380635261536 -0.2492087334394455 -0.8056405186653137 -0.9308571815490723 0.16308681666851044 0.13157081604003906 -0.0601432062685489
-0.08181028813123703 0.5495608448982239 0.3162820637226105 0.044546108692884445 0.7944148182868958 0.3056371808052063 -0.31843802332878113 -0.1738629937171936 0.156353
47095 -0.4021967947483063 1.411456823348999 -0.4098291695117950 4 0.3186410367488861 -1.1165008544921875 -0.05547158420085907 0.7745139002799988 0.43387165665626526 -0.2
.2594403326511383 -1.0255253314971924 0.7162691950798035 -0.3134252727031708 -0.33627140522003174 0.20484042167663574 -0.7187500596046448 -0.9545416235923767 -1.2280001
1.0184807777404785 0.5656768083572388 0.29735568165779114 -0.6946800947189331 -1.0299690961837769 1.2491761445999146 0.6057869791984558 0.5968462824821472 -0.0320619828
7 -1.5654470920562744 0.10546226799488068 -0.3829757869243622 -1.0525271892547607 0.10693001747131348 0.7864586114883423 0.5608898997306824 0.24788601696491241 0.526809
864502906799 -0.3442825973033905 0.27733975648880005 -0.7152450084686279 -0.24689675867557526 -0.7228854298591614 -1.2055928707122803 0.0033024270087480545 1.1218823194
32605 -1.6373316049575806 1.6440918445587158 -0.43068304657936096 -0.4733005762100022 -1.275132298469543 5 -0.25246310234069824 1.0536020994186401 -0.9760722517967224 -0.0
2385823726654053 -0.4132073223590851 0.01643933542072773 0.2160707563616185 0.18427337706089002 0.47540250420570374 -0.5665438771247864 -0.25542807579040053 1.06030929088
62297058105 0.017322789877653122 -0.11623325198888779 -0.7282307147979736 -0.9285045862197876 -0.8524077534675598 0.11137823760509491 1.0579420328140259 0.4089306890964
97026824951 -0.03119588829576969 0.10418453812599182 0.5999304056167603 -1.0119234323501587 -0.3879777193069458 -0.2799862325191498 -1.4234467744827227 0.108614973723888
-0.2590559422969818 0.5961945652961731 0.47084879875183105 0.4812585413455963 -1.0473358631134033 0.00704399636015296 0.5069509744644165 0.07217913866043091 0.44964727
28891 -0.5169460773468018 -1.1743741035461426 -0.7116754651069641 -0.4780204296112060 5 -1.6063034534454346 -0.16897273063659668 1.5204949378967285 -0.1661432683467865 -0
80347061157 0.4028494656085968 -0.18535855412483215 0.7548332810401917 0.3123325705528259 0.6593093872070312 0.3685827851295471 0.21299351751804352 0.9169759750366211
04139697551727 -0.29513370990753174 0.5013274550437927 -0.21574245393276215 -0.5136917233467102 -0.9796431064605713 0.28284478187561035 0.15777684718036652 0.992968142
of 0.2614080011844635 -0.05700542405247688 0.050627123564481735 -0.12816961109638214 -0.5257653594017029 0.06948389858007431 0.34602680802345276 0.34332260489463
928033351898 -0.11494740843772888 0.4615645110607147 0.34988754987716675 0.1981586515903473 0.06026952341198921 -0.8913010954856873 0.10875722020864487 0.09371425211142
095 -0.10814546048641205 -0.2949752211570740 0.4853720366954803 5 -0.7861011624336243 0.3077158033847809 -0.875210702419281 0.5334144234657288 0.8603790998458862 0.183168
0.014857389964163303 -0.19131046533584595 -0.14025750756263733 0.22238968312740326 -0.11352309584617615 0.12927047908306122 -0.2540449798107147 0.40016505122184753 -0.0
76 0.034201633318514824 -0.4061053991317749 -0.03724380955100059 5 0.3396824598312378 0.5161757469177246 -0.4832783043384552 -0.5008599162101746 0.5406717061996646 0.06227
-0.5704774260520935 -0.10188032686710358 -0.5763516426086426 -0.8384292721748352 -0.2848418653011322 0.05828262493043356 -0.07143948227167135 0.31795915961265564 -0.1265
42 0.23032259941101074 0.2935047149658203 0.11147939413785934 0.06336495280265808 0.1887342184782028 2 0.16351547837257385 -0.36865437030792236 -0.1314295083284378 -0.53
-0.7546107769012451 0.06165650486946106 0.47786179184913635 -0.0876844599843025 2 0.32733067870140076 0.9853317141532898 -0.045827437192201614 -0.12225066870450974 -0.56
74774 0.058116666972637177 -0.0871891975402832 -0.4840514063835144 -0.21969972550868988 -0.00875602662563324 -0.026952695101499557 0.24102148413658142 0.023833980783820
4354038239 0.11455033719539642 -0.20366010069847107 -0.18621009588241577 -0.28095537424087524 -0.07885187864303589 0.11217619478702545 -0.1198195070028305 -0.4577892720
01382040977478 0.0042733740992844105 -0.5538504123687744 0.4073474705219269 0.33703216910362244 -0.38002899289131165 -0.6938464045524597 -0.011663113720715046 -0.09683
3527934551239 -0.30656880140304565 -0.05122643709182739 0.46377986669540405 0.1711607426404953 0.3399849832057953 0.4210023581981659 -0.7595409750938416 -0.04228069633
07845306396 0.32652902603149414 -0.1293913573026657 0.46423065662384033 0.2933784425258636 5 -0.24779483675956726 -0.033143870532512665 0.15377464890480042 0.3256935179
172455 0.044297799468040466 0.45039498805999756 -0.2799057066440582 -0.2301137000322342 0.16503505406763885 0.11548270285129547 -0.07973808050155564 0.4481073021888733 -
4158 -0.3439595401287079 -0.14692458510398865 0.45710164308547974 -0.19501779973506927 0.30751246213912964 -0.14463791251182556 -0.25958913564682007 -0.2648999691009521
2010193 -0.3682197034358978 0.29388511180877686 -0.007812407799065113

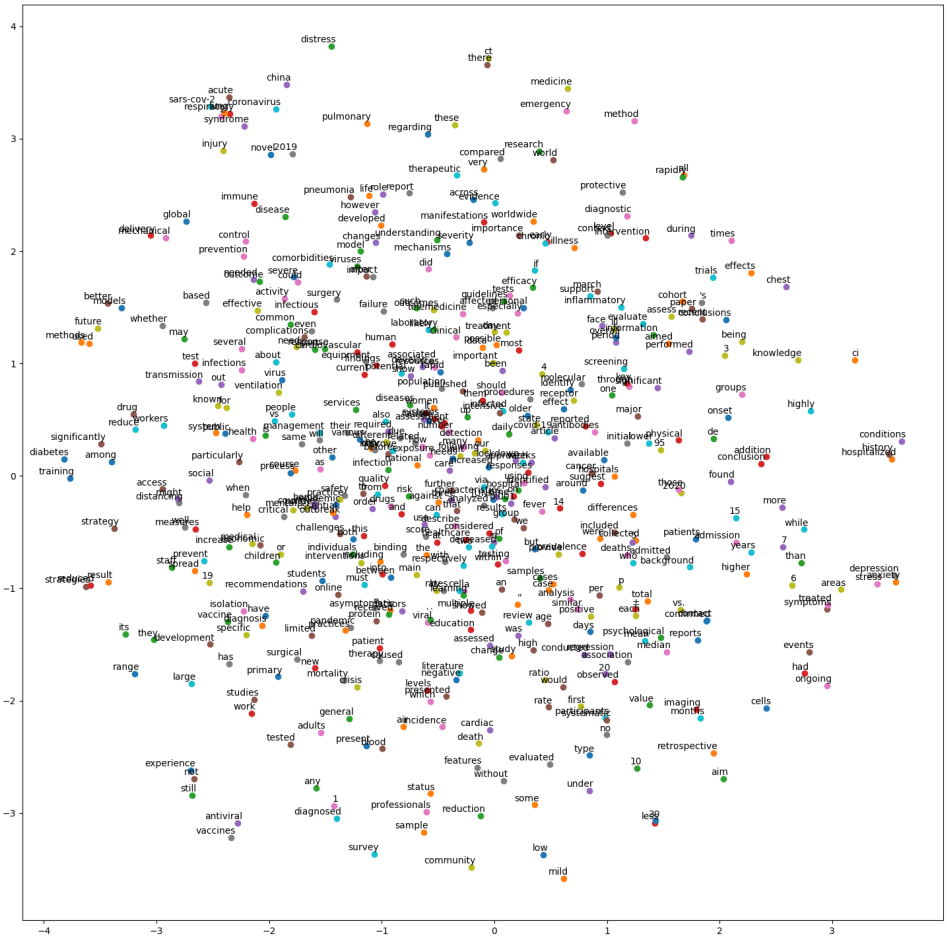
BERT Visualization



Word2Vec Vectors

[UNK] -0.04881330579519272 0.07739069312810898 -0.056522533929753876 -0.090730577770729065 -0.11317484825849533 0.05839929729700885 0.022058576345443726 0.04351604357
41718006134 -0.011797694489359856 0.10310874134302139 0.011879590339958668 0.116052135825157 17 -0.006655462551862001 -0.0010937526822090 15 -0.04330721125006676 -0.08714
611858368
the -0.14333058893680573 -0.1587565541267395 0.14329899847507477 -0.3877052068710327 -0.10853473842144012 -0.030758662149310112 0.04662489891052246 -0.024330385029
78235 -0.04327136278152466 -0.07235678285360336 -0.3090354800224304 0.12058231979608536 -0.020435530692338943 0.10384968668222427 0.28311672806739807 0.135265320539474
966949463
of 0.030921496450901 03 0.06394444406032562 0.2205295264720916 7 0.12312621623277664 0.13519074022769928 -0.33134761452674866 -0.08193377405405045 0.171484410762781
0996109992265701 0.08987542241811752 -0.06514536589384079 0.0856313705444336 -0.06625906378030777 0.11863439530134201 -0.18847365677735672 -0.07212437689304352 0.09053
7
and -0.022921787574887276 -0.27186140418052673 -0.13821113109588623 0.2388809472322464 -0.006997520104050636 0.015737395733594894 -0.2610098421573639 -0.0031904140
710476875305 0.14956840872764587 -0.26795533299446106 -0.29544875025749207 0.18868643045425415 -0.15007612109184265 0.13892631232738495 -0.015536155551671982 0.1629645
803672458976507
in 0.07111769169569016 -0.09316769242286682 -0.024860944598913193 0.12261395901441574 -0.32726961374282837 0.15798386931419373 -0.02231913059949875 -0.03950957953
0189832113683223 72 -0.16765150427818298 -0.17081472277641296 0.033556707203388214 -0.23027724027633667 0.10600022226572037 0.17270736396312714 0.2785123884677887 -0.06
to -0.07181657850742 34 -0.06219632923603058 -0.11162675172090 53 -0.035193767398595 81 0.40334665775299 07 0.009930439293384552 -0.06467465311288834 -0.49054917693133
26825 0.08831381797790527 0.021515537053346634 -0.434296190738678 -0.2705412209033966 -0.0353405736386776 -0.036785200238227844 -0.09575325250625 61 0.13779638707637787
a -0.25891706347465515 -0.2228064388036728 0.11413641273975372 -0.333120733499527 0.05918675661087036 -0.23488295078277588 0.27852168679237366 0.05177316442131995
17957484722137 0.24900034070014954 0.11283621191978455 -0.12552566826343536 -0.07017210125923157 0.009417561814188957 0.03100868873298168 2 0.07272831350564957 0.141228
with -0.09855984151363373 -0.23852476477622986 0.11758941411972046 0.06500956416130066 -0.06871606409549713 -0.2874845266342163 -0.08065295964479446 0.1696904450654
16490524634718895 0.21965739130973816 -0.3124369978904724 -0.10290412604808807 -0.04615779685883713 0.040104590356349945 0.21833129227161407 0.4410780966281891 0.22014
for -0.08885828405618668 0.04963880032300949 -0.03774642199277878 -0.06492927670478821 0.4045398235321045 -0.11071208864450455 -0.04672267660498619 0.1040973588824
1584 -0.21206845343112946 -0.039591915905475616 -0.626276969909668 -0.21396154165267944 0.06573865562677383 -0.08018310368061066 0.13445967435836792 0.12017136067152023
covid-19 0.33330416679382324 0.07484094798564911 -0.14706000068569 1833 -0.022302521392703056 -0.19168257771331787 -0.11672676354646683 0.08991703391075134 0.14642255351810
55426 -0.04511909931898117 0.2939341962337494 0.24353934824466705 0.17898909747600555 -0.009478721767663956 -0.3241097927093506 -0.07146181166172028 -0.1966229081153866
patients 0.04939527064561844 -0.14423398673534393 -0.10975630581378937 -0.25172340869903564 -0.08555308729410172 -0.40732452273368835 0.1958676278591156 0.3668800294399
92216610908 5 -0.41730549931526184 0.4436086118221283 -0.021435552526831627 0.037291403859853745 0.03907901048660278 -0.09325745701789856 -0.05411481112241745 0.17745845
16
were 0.5449550747871399 0.12638379633426666 0.037353437393907 3 0.08160431683063507 -0.11231604963541031 0.1520289033651352 -0.36640504002571106 0.10934747010469437
883780777454376 -0.6140636205673218 -0.09264766424894333 0.023603010922670364 -0.13203932344913483 0.09088529646396637 0.24826601147651672 0.026199063286185265 0.11704
2937
is -0.2018497884273529 -0.014681762084364891 -0.27701535820961 -0.3487420082092285 0.5759435892105103 0.1453063189983368 0.08596295118331909 -0.42484018206596375
2977199554443 0.04105306416749954 -0.10346832126379013 0.026820257306098938 -0.1885482668876648 0.4364898204803467 -0.06917911767959595 -0.06044343486428261 0.52001386

Word2Vec Visualization



Discussion

In these two graphs, we can find that in BERT result word mainly separate in two parts while Word2Vec result stay in one part. Also, in BERT result, there exits some nodes which stay very close. Although, Word2Vec result's nodes all exist in one part, the distance between two nodes is longer than BERT model. BERT result is more obvious, which may be one reason of playing a better role in word embedding.