Word Embedding by Using Word2Vec and BERT

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

Author: Kewei Zhao
Date: 2021-5-22

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

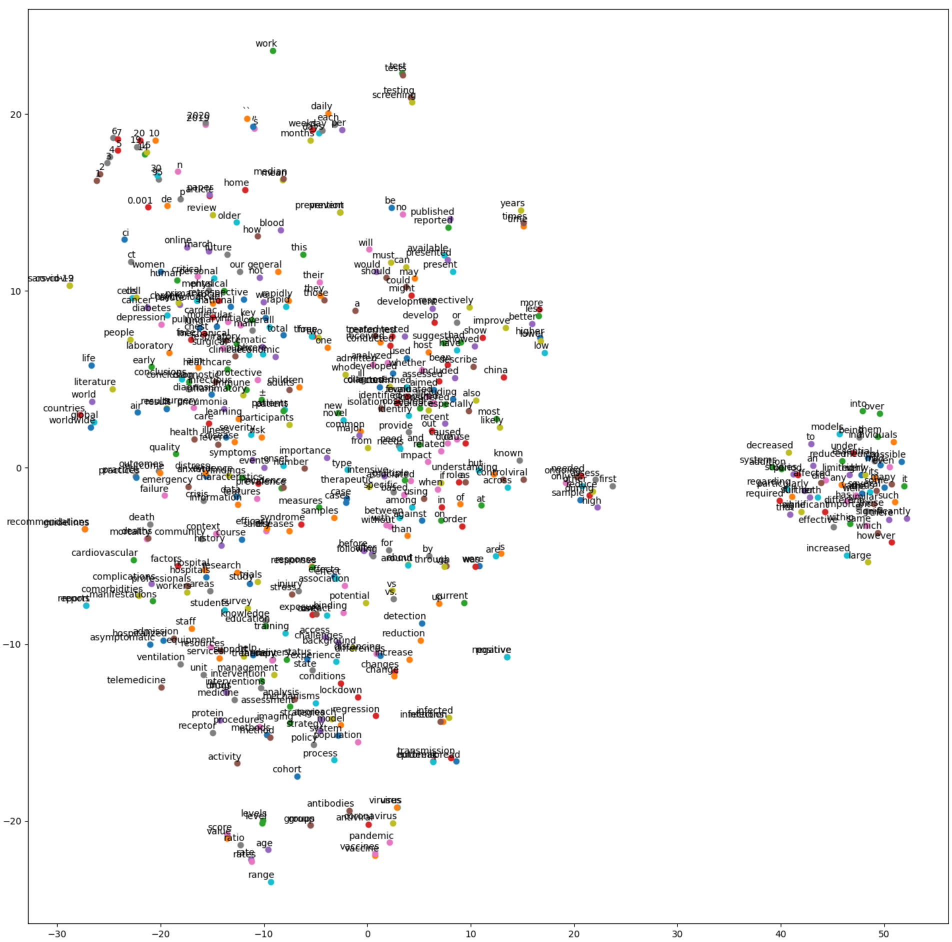
```
python3 Bert_4_Litcovid_WordEmbedding.py
```

Result

BERT Vectors

the 0.35628464818000793 -0.5250380635261536 -0.2492087334394455 -0.8056405186653137 -0.9308571815490723 0.16308681666851044 0.13157081604003906 -0.060143206268548
-0.08181028813123703 0.5495608448982239 0.3162820637226105 0.044546108692884445 0.7944148182868958 0.3056371808052063 -0.31843802332878113 -0.1738629937171936 0.156353
47095 -0.4021967947483063 1.411456823348999 -0.40982916951179504 0.3186410367488861 -1.1165008544921875 -0.05547158420085907 0.774513900279988 0.4338716566562652 -0.2
.2594403326511383 -1.0255253314971924 0.716269195079803 -0.3134252727031708 -0.33627140522003174 0.20484042167663574 -0.7187500596046448 -0.9545416235923767 -1.228000
1.0184807777404785 0.5656768083572388 0.29735568165779114 -0.694680094718933 -1.0299690961837769 1.2491761445999146 0.6057869791984558 0.5968462824821472 -0.032061982
7 -1.5654470920562744 0.10546226799488068 -0.3829757869243622 -1.0525271892547607 0.10693001747131348 0.7864586114883423 0.5608898997306824 0.2478860169649124 0.526809
864502906799 -0.3442825973033905 0.27733975648880005 -0.7152450084686279 -0.24689675867557526 -0.7228854298591614 -1.2055928707122803 0.0033024270087480545 1.12188231
32605 -1.6373316049575806 1.6440918445587158 -0.43068304657936096 -0.473300576210022 -1.2751322984695435 -0.25246310234069824 1.053602099418640 -0.9760722517967224 -0.0
2385823726654053 -0.4132073223590851 0.01643933542072773 0.216070756316185 0.1842733770608902 0.47540250420570374 -0.5665438771247864 -0.2554280757904053 1.06030929088
62297058105 0.017322789877653122 -0.11623325198888779 -0.7282307147979736 -0.9285045862197876 -0.8524077534675598 0.11137823760509491 1.0579420328140259 0.40893068909
97026824951 -0.03119588829576969 0.10418453812599182 0.5999304056167603 -1.0119234323501587 -0.3879777193069458 -0.2799862325191498 -1.423446774482727 0.10861497372388
-0.2590559422969818 0.5961945652961731 0.47084879875183105 0.4812585413455963 -1.0473358631134033 0.00704399636015296 0.506950974464416 0.07217913866043091 0.4496472
28891 -0.5169460773468018 -1.1743741035461426 -0.7116754651069641 -0.47802042961120605 -1.6063034534454346 -0.16897273063659668 1.5204949378967285 -0.1661432683467865 -
80347061157 0.402849465608596 -0.18535855412483215 0.7548332810401917 0.3123325705528259 0.6593093872070312 0.3685827851295471 0.21299351751804352 0.916975975036621
04139697551727 -0.29513370990753174 0.5013274550437927 -0.21574245393276215 -0.5136917233467102 -0.9796431064605713 0.2828447818756103 0.15776784718036652 0.99296814
of 0.2614080011844635 -0.05700542405247688 0.05062712356448173 -0.12816961109638214 -0.5257653594017029 0.06948389858007431 0.3460268080234527 0.3433226048946
928033351898 -0.11494740843772888 0.4615645110607147 0.34988754987716675 0.1981586515903473 0.06026952341198921 -0.8913010954856873 0.10875722020864487 0.0937142521142
095 -0.10814546048641205 -0.294975221157074 0.48537203669548035 -0.7861011624336243 0.3077158033847809 -0.875210702419281 0.5334144234657288 0.8603790998458862 0.18316
0.014857389964163303 -0.19131046533584595 -0.14025750756263733 0.22238968312740326 -0.11352309584617615 0.12927047908306122 -0.2540449798107147 0.4001650512218475 -0.0
76 0.03420163318514824 -0.4061053991317749 -0.037243809551000595 0.3396824598312378 0.5161757469177246 -0.4832783043384552 -0.5008599162101746 0.54067170619964 0.06227
-0.570477426052093 -0.1018803268671035 -0.5763516426086426 -0.8384292721748352 -0.2848418653011322 0.05828262493014335 -0.071439482271671 0.31795915961265564 -0.126
42 0.23032259941101074 0.29350471496582 0.11147939413785934 0.06336495280265808 0.18873421847820282 0.16351547837257385 -0.36865437030792236 -0.1314295083284378 -0.53
-0.754610776901245 0.06165650486946106 0.47786179184913635 -0.08768445998430252 0.32733067870140076 0.9853317141532898 -0.045827437192201614 -0.12225066870450974 -0.56
74774 0.058116666972637177 -0.0871891975402832 -0.4840514063835144 -0.21969972550868988 -0.008756026625633 -0.026952695101499557 0.2410214841365814 0.02383398078382
4354038239 0.11455033719539642 -0.20366010069847107 -0.18621009588241577 -0.28095537424087524 -0.07885187864303589 0.11217619478702545 -0.119819507002830 -0.45778927
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 -0.759540975093841 -0.04228069633
07845306396 0.32652902603149414 -0.1293913573026657 0.46423065662384033 0.29337844252586365 -0.24779483675956726 -0.03314387053251266 0.1537746489048004 0.3256935179
172455 0.044297799468040466 0.4503949880599975 -0.2799057066440582 -0.2301137000322342 0.16503505408763885 0.11548270285129547 -0.0797380805015564 0.4481073021888733 -
4158 -0.343959540128707 -0.14692458510398865 0.45710164308547974 -0.19501779973506927 0.30751246213912964 -0.14463791251182556 -0.2595891356468200 -0.264899969100952
2010193 -0.3682197034358978 0.29388511180877686 -0.007812407799065113

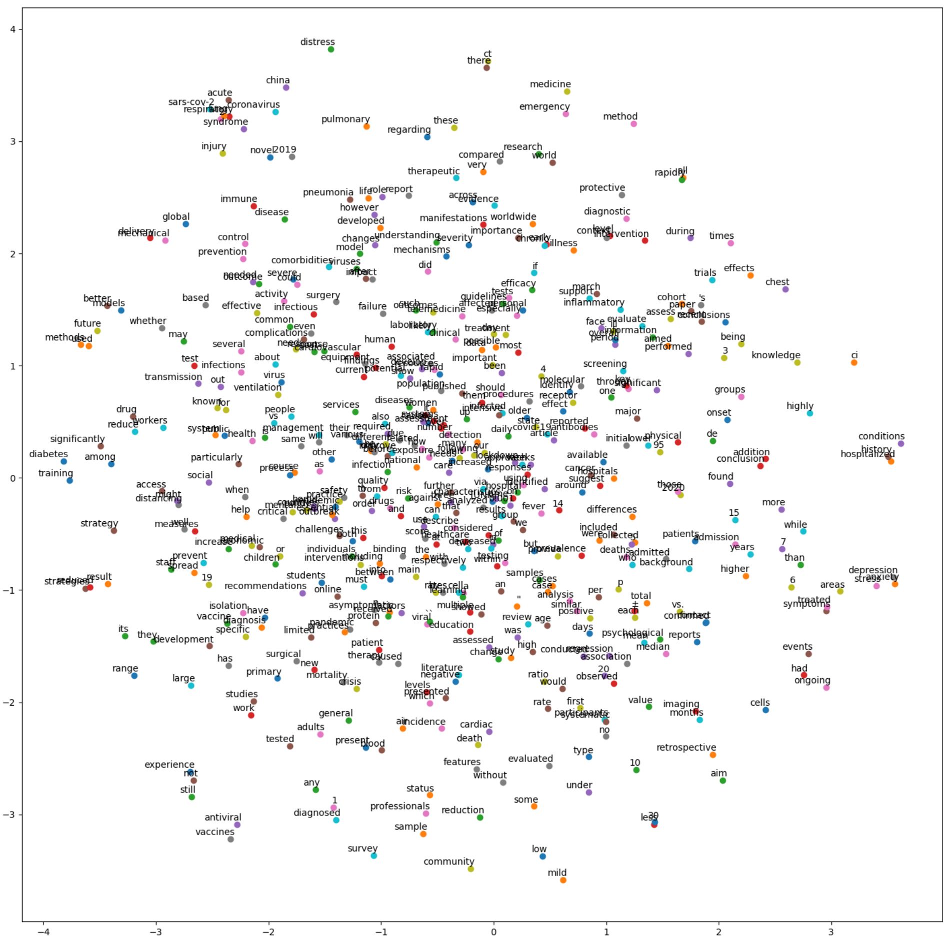
BERT Visualization



Word2Vec Vectors

[UNK] -0.04881330579519272 0.07739069312810898 -0.05652253329753876 -0.09073057770729065 -0.11317484825849533 0.05839929729700085 0.022058576345443726 0.0435160435
41718006134 -0.011797694489359856 0.10310874134302139 0.011879590339958668 0.11605213582515717 -0.006655462551862001 -0.001093752682209015 -0.04330721125006676 -0.0871
611858368
the -0.14333058893680573 -0.1587565541267395 0.14329899847507477 -0.3877052068710327 -0.10853473842144012 -0.030758662149310112 0.0466248989105224 6 -0.02433038502
78235 -0.04327136278152466 -0.07235678285360336 -0.3090354800224304 0.12058231979608536 -0.020435530692338943 0.10384968668222427 0.2831167280673980 7 0.135265320539474
966949463
of 0.03092149645090103 0.06394444406032562 0.22052952647209167 0.12312621623277664 0.13519074022769928 -0.33134761452674866 -0.08193377405405045 0.1714844107627
10996109992265701 0.08987542241811752 -0.06514536589384079 0.0856313705444336 -0.06625906378030777 0.11863439530134201 -0.1884736567735672 -0.07212437689304352 0.09053
7
and -0.022921787574887276 -0.27186140418052673 -0.13821113109588623 0.2388809472322464 -0.006997520104050636 0.015737395733594894 -0.2610098421573639 -0.003190414
710476875305 0.1495684087276458 7 -0.26795533299446106 -0.29544875025749207 0.18868643045425415 -0.15007612109184265 0.13892631232738495 -0.01553615555167198 2 0.162964
803672458976507
in 0.07111769169569016 -0.09316769242286682 -0.024860944598913193 0.12261395901441574 -0.32726961374282837 0.15798386931419373 -0.0223191305994987 5 -0.0395095795
018983211368322372 -0.16765150427818298 -0.17081472277641296 0.033556707203388214 -0.23027724027633667 0.10600022226572037 0.17270736396312714 0.2785123884677887 -0.06
to -0.0718165785074234 -0.06219632923603058 -0.1116267517209053 -0.03519376739859581 0.4033466577529907 0.009930439293384552 -0.0646746531128883 4 -0.490549176931
26825 0.08831381797790527 0.021515537053346634 -0.434296190738678 -0.2705412209033966 -0.0353405736386776 -0.036785200238227844 -0.0957532525062561 0.13779638707637787
a -0.25891706347465515 -0.22280643880367 28 0.11413641273975372 -0.333120733499527 0.05918675661087036 -0.23488295078277588 0.27852168679237366 0.051773164421319
17957484722137 0.24900034070014954 0.11283621191978455 -0.12552566826343536 -0.07017210125923157 0.009417561814188957 0.031008688732981682 0.07272831350564957 0.14122
with -0.09855984151363373 -0.23852476477622986 0.11758941411972046 0.06500956416130066 -0.06871606409549713 -0.287484526634216 3 -0.08065295964479446 0.16969044506 54
16490524634718895 0.21965739130973816 -0.3124369978904724 -0.10290412604808807 -0.04615776985883713 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 0.120171360671502
covid-19 0.33330416679382324 0.07484094798564911 -0.14706000685691833 -0.022302521392703056 -0.1916825771331787 -0.11672676354646683 0.08991703391075134 0.146422535181
55426 -0.04511909931898117 0.2939341962337494 0.24353934824466705 0.17898909747600555 -0.009478721767663956 -0.3241097927093506 -0.07146181166172028 -0.19662290811538
patients 0.04939527064561844 -0.14423398673534393 -0.10975630581378937 -0.25172340869903564 -0.08555308729410172 -0.40732452273368835 0.195867627859115 6 0.3668800294399
922166109085 -0.41730549931526184 0.4436086118221283 -0.021435525268316 27 0.037291403859853745 0.03907901048660278 -0.09325745701789856 -0.05411481112241745 0.1774584
16
were 0.5449550747871399 0.12638379633426666 0.037353437393903 73 0.08160431683063507 -0.11231604963541031 0.1520289033651352 -0.3664050400257110 6 0.109347470104694
883780777454376 -0.614063620567321 8 -0.09264766424894333 0.02360301092267036 4 -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 0.5200138

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.