

Nearest Neighbor Interpretability of Transformer-Based Model Embeddings

Nathan Choi Kenneth Lin Simone Ong

School of Information, University of California Berkeley
{nathan.choi, kennyalexlin, simoneong}@berkeley.edu

Abstract

We study the extent to which modern transformer-based architectures can be used to track linguistic usage change across corpora. Specifically, we pre-train 7 BERT models on different corpora using the masked language modeling (MLM) training task and examine the resulting word embeddings within the shared vocabulary space between pairs of models. We apply a simple framework for identifying usage change through the analysis of these contextualized word embeddings, which is inspired by and compared to previously attempted methods (Word2Vec) that have not utilized contextual embeddings during training. We originally intended to show that by slightly altering the process to leverage a model that incorporates self-attention, we could achieve similar, if not better results while keeping the rest of the usage change analysis the same. Our results imply that BERT models may be less suitable for this task than classical approaches such as Word2Vec, although we note that resource constraints may have limited the performance of our models.

1 Introduction

Since their introduction in 2017, transformer-based models have achieved state-of-the-art performance on a variety of natural language processing tasks. Their success is widely due to the introduction of positional encoding and self-attention, which enables them to augment token-level embeddings with information about their surrounding context. We seek to understand how word embeddings produced by transformer models differ from embeddings produced by classical models, particularly with respect to interpretability.

Due to the evolution of mediums from which people obtain news and information, it is worth being able to study this generational change as language develops over time. The most straightforward way to do so is by looking across multiple

sources over time and how the context of language may have changed. There are many emerging opportunities for representing language with word embeddings with the development of many new language models (BERT in our project’s case). These can provide important insights on the current state of Natural Language Processing, especially when compared to past methods and algorithms of word embedding interpretability.

2 Background

Our work is inspired by an earlier implementation that resulted in a stable method for detecting changes in word usage across different corpora (Gonen et al., 2020). The methodology takes two separate corpora (with similar vocabularies but a difference in one aspect, e.g. time frame) as input and returns a ranked list of words that are sorted from being the most likely to have changed to the least likely. This ranking is produced by using Word2Vec to create word vectors for each of the corpora, then examining the 10 Nearest Neighbors of each word in their respective vector spaces. The words with the least amount of intersecting nearest neighbors are deemed to have undergone the most usage change. The position of the words is not prioritized, but rather the presence of the word itself and what words are surrounding it to create its context. The way that (Gonen et al., 2020) compared their results was by using their nearest neighbor ranked list compared to the AlignCos method (Hamilton et al., 2016). The AlignCos method trains the word embeddings, aligns the two vector spaces, and then calculates a cosine distance between the two points in which that word is placed inside the two spaces. If there is a larger distance between the two corpora’s spaces, it implies a larger change in usage.

Word embeddings and contextual embeddings share the same goal of creating a vector represen-

tation of a word. However, they do so in different ways which can inherently change the nature of the resulting word vectors. Word embeddings focus on the definition of the word rather than the context around the word to learn its numerical vector input. This can lead to a loss of deeper meaning for words that could have multiple meanings such as "bank", which would only be given one definition. Contextualized word embeddings help build a word vector that is based on its context, which is then able to capture the correct definition of the word as it's used in that specific context.

BERT models utilize positional embeddings and actually looks both before and after a given token to learn context to be able to train a bidirectional Transformer (Devlin et al., 2019). The value of being able to have a multi-headed self-attention to pick up context is that it looks at different parts of the input at the same time which gives multiple perspectives of context. This helps to enhance the learning of complex word relationships. This is unlike Word2Vec, which uses shallow neural networks with two main models, Continuous Bag of Words and Skip-gram. Continuous Bag of Words looks at the inputs taken in by the neural net and then predicts the target word as close as possible to the context it was given. Skip-gram uses the current word as the input and then predicts words that are close in terms of meaning. (Turing, 2024)

Considering (Gonen et al., 2020) and (Devlin et al., 2019) together, we were interested in "combining" the two methodologies with this more updated BERT model in order to add context to "improve" the embedding ability from (Gonen et al., 2020). In theory, this would create different results, as the training now utilizes contextual embeddings; so there is reason to believe that there could potentially be some sense of improvement. This would of course be up to subjective interpretation.

3 Methodology

Our data comes from the Celebrity Profiling corpus (Wiegmann et al., 2019) which is a collection of celebrity tweets that has additional identifiers including age, gender, occupation, etc. We are utilizing the same splits and datasets as the original paper (Gonen et al., 2020), which look at age: young (celebrity birth year 1990-2009) and old (celebrity birth year 1950-1969), gender: male and female, and pairwise splits for occupation: performer and sports, performer and creator, creator and sports.



AGE (YOUNG VS. OLDER)	
NN	neighbors in each corpus
dem	dese, yuh, them, nuh, dey, ayye, dats, tha, betta, fuk repub, democrats, centrist, manchin, primaries, party's, alp, dfl, gopers, repubs
dam	damm, mannnnn, mannn, mane, huh, ahb, oo, buggin, koo, mannn dams, basin, river, dredging, reservoir, drainage, wastewater, sewerage, refinery, canal
rep	reppin, wear, allegiance, all-american, wildcat, alumni, tryout, hoosier, recruit, ua sen., congresswoman, chairwoman, co-chairs, gazelka, salazar, amb, comptroller, staffer, cong
assist	points, shutout, scoresheet, scored, pts, hatrick, sheet, nil, sacks,  assisting, contact, coordinate, locating, coordinating, administer, equip, consular, deploy, locate
pr	cameron, -pr, erik, lap, sergeant, laps, tundra, teamjev, caution, restart stunt, puerto, promotional, rico, creative, ploy, hire, spin, freelance, fema
fr	frfr, forreal, foreal, lmaooo, madd, tho, bck, bruhh, lmao, fwm pavone, incl, from, wrk, ger, joseph, covey, env, w, ans
joint	jawn, fusion, scorpion, sumn, spot, db, cb, joints, mgmt, fye high-level, convened, minsk, two-day, bilateral, counter-terrorism, delegations, asean, convene, liaison
mega	 fantastic, simulator, macau, lotus, fuji, bmw, awesome, mclaren, fab gujarat, becos, multi-billion, gta, rupees, dollar, maharashtra, major, crores, multi-million
flow	beard, vibin, jeezy, drizzy, lite, mohawk, dreads, sauna, boomn, vibe illicit, influx, accumulation, moisture, absorb, overwhelm, heart's, drains, curtail, diverting
icymi	superintendent, bureau, commissioner, spokesman, exec, state's, prosecutor, Reuters, montgomery, Conway re-upping, reichert, newsmakers, sherrod, column, arizona's, otl, holcomb, rundown, wrap-up

Figure 1: Sample word comparison results from the original (Gonen et al., 2020) paper. This shows the top 10 most changed words in their respective corpora and their respective nearest neighbors

We used the same labeling algorithm that (Gonen et al., 2020) produced in order to ensure consistency between the datasets while varying our embeddings. From here, we deviate from the original paper to build upon its pipeline by applying a WordPiece tokenizer to the celebrity tweet data to eventually train our own BERT-style model instead of their Word2Vec. We train this BERT model with the Masked Language Modeling task on each of the aforementioned corpora (Age: Young, Age: Old, Gender: Male, Gender: Female, Occupation: Performer, Occupation: Sports, and Occupation: Creator).

Given that our project is inspired by previous works, our baseline is the Word2Vec word usage change comparisons from the original paper that inspired our project. While it can be difficult to judge a comparison of how the usage of a word could change over time from the outcome of the word embeddings alone, the work from (Gonen et al., 2020) produced very compelling results in our opinion at identifying words that have changed in use across the different corpora. A sample of their results is shown in Figure 1, which shows multiple words that have undergone usage change between younger and older celebrities. It was therefore our goal to replicate their work using a BERT model to implement a transformer that could learn with context. Our process began by first processing the tweets, as we created and trained a BERT word-piece tokenizer from scratch. We then performed a test/train split on our dataset and trained our BERT model. Once the BERT model had been trained, our goal was then to analyze our model's embeddings to evaluate interpretability and potential usage change. To compare our models with

the original Word2Vec model from the original paper, we looked at the same 10 words that (Gonen et al., 2020) used and looked at the K-Nearest Neighbors within the vector space of our models. After many trials, our 7 BERT models (one for each corpus) used a vocab size of 30522 and were each trained on 1,280,000 tweets with a similarly sized test set. We do an additional validation on a pre-trained BERT model to compare our models' results against a fully trained BERT model (BERT Base Uncased).

4 Results and Discussion

Due to the exploratory nature of our experiments to evaluate word embedding interpretability, it is difficult to specifically quantify our models' ability to produce interpretable word embeddings that could then be used to find usage change. The full scope of our results compared to the original paper's corpus pairings can be found in the Appendix A. While the development of this project proved to be a good learning experience for ourselves, our BERT models demonstrated a very poor ability to create intuitive word embeddings. Using our version of Figure 1 shown below in Appendix A.2 as a comparison, it is apparent that our models for both the "Young" and "Old" corpora struggled to produce interpretable embeddings. More specifically, the K-Nearest Neighbors for each of the words are not similar at all to the word being evaluated. For example, the word "assist" has an assortment of words like "healed", "bruin", and "misleading" as neighbors for the Young age group celebrities. The Old age group nearest neighbors do not perform any better with more random words like "logs", "essays", and "world". One key difference that we expected and did see in our results was a slight difference in word tokens, due to the different tokenization schemes between Word2Vec from the original paper and BERT Tokenization. Word2Vec uses entire words while BERT uses WordPiece, so in some instances we only see parts of words in our results like "maur", "soci", etc.

Although they are disappointing, the poor results can most likely be attributed to multiple factors. Our biggest challenge as a team was overcoming the immense computational load of training a BERT model on millions of tweets. Our models each took about 4 hours to train and another 2 hours or so to extract the embeddings from them. Furthermore, our vocab size was limited compared

to the (Gonen et al., 2020) paper at 30556 words compared to their vocab sizes ranging from 42000 - 114000 words depending on the corpora. Our local machines and Google Colab instances could only handle smaller batches of tweets before running out of memory, so our final model parameters needed to balance having enough data while not training for too long. Unfortunately, our models' seeming inability to produce meaningful relations between its K-Nearest Neighbors lead us to believe that these may not be suitable for analysis from a word usage change perspective. However, our poor results can still be discussed and compared to the original paper.

To further validate the efficacy (or lack thereof) of our BERT-style models, we examined the K-Nearest Neighbors themselves to see if they were even close to the actual words within the vector space. Although the nearest neighbors were obviously not very intuitive for our interpretation, we could leverage cosine distance to evaluate relatively how close each of the neighbors were to the word at hand. We compared our models against the official BERT Base Uncased model to have a source of truth. We took the average cosine distance of the 10 Nearest Neighbors of both models for the words "basketball", "track", and "wing". The results are shown below:

Cosine Distance Comparison	
Model	Average Cosine Distance
BERT Base Uncased	0.2688
"Occupation: Sports"	0.9468

From the above table, it is evident that the nearest neighbors from our BERT models are not very close to the word at all. For reference, the nearest neighbors of the BERT Base Uncased model for the three analyzed words are shown in Appendix A.6. It is clear that with the fully trained BERT Base Uncased model, the nearest neighbors are much more similar to the given word than with our models and provide a much more ideal example of what we would have liked ours to produce. These results are much more comparable to those of the original paper's as well.

It is very apparent that our BERT-style models did not reach the full potential of the Word2Vec model from (Gonen et al., 2020) or the BERT Base Uncased model. This puts into perspective how important a large training sample is for a model to be able to extract meaning from large bodies of text. Our training scope is minuscule in comparison to

that of the BERT Base Uncased model. If we had more compute power we would have been able to train on more data and also could have explored adding the Next Sentence Prediction training task to the training pipeline as well. This could be an improvement for the future to get more insightful results.

5 Conclusion

In our overall goal of swapping word embeddings with contextual embeddings to see if word context changes over time, our results were slightly inconclusive, but we were still able to show that the process was replicable with the ability to swap out one tokenizer and modeling method. There were definitely additional measures we could have looked into or more data we could have trained on to have a more robust outcome; but with the current resource limits, this was not attainable. Given more compute resources and training time, this exploratory word embedding analysis can still have potential in producing interpretable word embeddings that can then be used to evaluate word usage change across corpora.

References

- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Hila Gonen, Ganesh Jawahar, Djamé Seddah, and Yoav Goldberg. 2020. [Simple, interpretable and stable method for detecting words with usage change across corpora](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 538–555, Online. Association for Computational Linguistics.
- William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. [Diachronic word embeddings reveal statistical laws of semantic change](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.
- Inc. Turing. 2024. Guide on word embeddings in nlp. <https://www.turing.com/kb/guide-on-word-embeddings-in-nlp>.

[guide-on-word-embeddings-in-nlp](#). Accessed: 2024-07-31.

A Appendix: Results

These are the results for our final models' embeddings for words to be compared to the original paper from (Gonen et al., 2020). In each table of entity1 vs. entity2, the entity1 K-Nearest Neighbors are listed on the first row, while the entity2 K-Nearest Neighbors are listed on the second row

A.1 Gender: Female vs. Male

Gender: Female vs. Male	
Word	nearest neighbors in each corpus
bra	festive, conserv, elect, beach, niki, pep, surface, feelin, soci, calming flaw, himself, fav, recovered, goal, nintend, pink, cheated, -, cvs
clutch	conquer, woops, coyote, bjj, kimmy, lust, diy, marian, suggestinos, ironman frozen, saved, exhib, lengthy, ta, helsinki, violates, preferences, toul, wiped
mm	uplifting, mom, letters, stockings, resemblance, service, rhetor, funny, yummy, coolest jav, declines, muffin, stroll, relating, kristen, expands, legacy, barack, anniversary
mc	actress, noir, freel, whaaaaat, deter, artisan, neither, apocalyptic, buzz, waitin cool, sacramento, naa, valerie, initially, welter, hi, unique, budgets, roundup
gp	plugged, thur, organizations, tyrant, comunic, player, hysterically, bury, angie, grrr writ, premiere, question, sacramento, silverstone, l, lux, yoo, yup, skills
keeper	sketch, mixing, opportunity, buds, reacting, fold, fleeting, releases, opposing, intake rebs, ratio, boule, cattle, license, brav, float, ages, goal, badger
nd	finland, unbelievably, purchases, pt, underneath, systemic, hamstring, brigade, district, poll manage, preferences, employ, february, lbc, saul, katz, notices, missus, cement
hay	hippie, cron, bernie, serge, diabetes, guacamole, dug, yol, clive, flare suc, newcastle, offside, hogan, renee, reunion, change, rumor, within, standard
steph	kiss, rabbi, penguin, shoe, hershey, instruction, build, kudos, powerball, crowd nothin, prejud, persu, stuff, roundtable, batsman, condescending, gujar, basic, westside
echo	notified, masterclass, key, ex, guil, encounter, goddesses, tragedy, roaring, differently brody, misc, restore, islamabad, television, appeared, ferry, entertain, cord, fandom

A.2 Age: Young vs. Old

Age: Young vs. Old	
Word	nearest neighbors in each corpus
dem	maggie tick hopefully leigh recovery karan fasting fitt subtweeting panda collie, implants, enfield, finances, syfy, valley, hayes, aussies, apo, dip
dam	kissing, minute, medalist, clipping, consulting, millers, manny, homer, actors, ama inspirations, flashing, turkish, punching, virat, lauderdale, prior, reel, rescheduled, sag
rep	spokesperson, officiating, twd, repe, dep, convers, study, advice, recie, capacity defer, with, degrees, uc, induct, chang, abol, sak, harmful, warmed
assist	consulting, ep, healed, bruin, popping, misleading, raz, zen, ehh, revising logs, essays, traum, hiring, bastard, gos, world, nys, treasures, navigate
pr	sprinting, arrest, sponsoring, pjs, rental, jai, xxx, kitting, massachusetts, coffe onward, steyn, cattle, root, caffe, expressions, unlocking, miguel, atmos, goat
fr	loans, sands, femme, pilgrim, sett, juventus, strive, princ, actor, dak surat, mubar, mysterious, hitting, swim, grew, deceit, sunny, trib, sim
joint	poor, eject, glow, thangs, aussie, thailand, motorway, taps, teacher, mathem djokovic, soundtrack, consul, gg, coincidentally, evacuees, strongly, accused, glam, ami
mega	lass, cluster, kuz, cat, beth, athens, savvy, tink, invite, 4 cameraman, gogg, raleigh, iron, undert, outlet, diplomat, appointments, decline, matchup
flow	massachusetts, albion, supplying, makeup, wives, princ, omaha, pjs, repe, fines crook, catalan, carrie, bulletin, conspir, lauderdale, tunnels, county, hikes, maur
icymi	karma, leverage, raving, joyful, ribbon, arrangements, pics, chopping, locate, coles jah, cba, cds, flashing, agony, wolff, sp, adjusting, suspended, progress

A.3 Occupation: Performer vs. Sports

Occupation: Performer vs. Sports	
Word	nearest neighbors in each corpus
blues	loudest, fuse, undecided, telegraph, meditating, trucks, chall, scand, onto, daught depress, infringing, skysports, edm, muslim, spam, refreshed, ballarat, ceases, hall
cc	smoothies, coaches, agriculture, biography, outfit, scare, hurrah, xoxoxox, mentor, demon tomas, slider, falls, fors, congress, craft, punts, ala, circuit, squaw
dub	initial, malaysia, penguin, tights, intimid, burn, less, hmmmmmm, donny, lagged deck, fortun, swindon, freck, recreational, incoming, shuts, fanny, recent, carlisle
bra	taboo, brainer, dock, peaked, answ, organising, porsche, contest, obit, balloons headwind, sensitive, brilliance, andi, rts, unjust, spent, showed, god, bff
track	round, feel, funk, while, pondering, consistently, spawn, kennedy, playoff, commend wheelie, milks, bradenton, caa, rel, matth, aww, uae, nelly, trucking
wing	herbs, presser, access, harry, gatsby, treasure, lucy, buckets, grants, chen octagon, mind, hiccups, bez, brass, forced, neither, athletes, adventurous, premise
par	kennedy, encounter, tacky, mcl, ooh, attended, caldwell, pett, mutant, nice slogan, kilt, mediocre, ability, driven, mandatory, sb, physically, jumb, consultation
mo	louis, extremely, apologizes, shin, uuuu, woody, duets, hob, brother, infectious hospice, snot, rory, disclaim, colt, vets, grammy, pulls, mf, ex
ace	bourbon, aol, milestone, skateboard, spielberg, badger, subtle, believer, bothering, titanic planner, unjust, obtained, jordans, snor, eston, byrd, seren, gut, low
duo	hex, diplom, donnie, cub, time, rhetoric, ark, table, boring, supposed baseballs, verdict, tix, stol, outage, hold, accomplishments, skyfall, els, parsons

A.4 Occupation: Creator vs. Sports

Occupation: Creator vs. Sports	
Word	nearest neighbors in each corpus
cc	treat, eds, oc, baby, foxes, unexpl, delta, lava, pav, explosive tomas, slider, falls, fors, congress, craft, punts, ala, circuit, squaw
op	peeps, galley, precise, persuasive, hmmmmmm, silhouette, tree, relig, keen, communist zing, cedric, establish, goodmorning, alcoholic, inspir, behalf, punching, labs, deck
blues	monit, apparel, shuff, jord, potential, , dell, rodriguez, apologizes, gerrymand depress, infringing, skysports, edm, muslim, spam, refreshed, ballarat, ceases, hall
origin	difficulties, bett, allegation, shakers, dian, tyson, openly, promo, appreciated, lifetime mistakes, happens, egregious, cancer, killings, wer, fick, rallies, convers, indonesian
wing	hammers, cute, blades, workouts ev foremost swung cbcnn pauline energy octagon, mind, hiccups, bez, brass, forced, neither, athletes, adventurous, premise
weigh	desired, quiet, shortlisted, coff, nur, baths, ignorant, prepares, tromb, expanded poetic, natal, bland, dojo, badges, fors, cya, yuan, keegan, counting
worlds	johann, king, mamas, eternal, gamble, loans, sequel, kam, demonstrations, strategy rachel, representing, operation, holocaust, confront, stra, incoming, jess, skid, kard
sessions	rbs, judicial, peps, traumatized, clicks, cummings, aspen, concert, den, ohhh fumble, bare, minions, rate, pockets, fart, photographer, stared, hemp, bath
track	thiel, embarrassed, humbly, utmost, irreg, diana, achilles, stab, considerably, brut wheelie, milks, bradenton, caa, rel, matth, aww, uae, nelly, trucking
presents	thesis, gron, backlash, bangs, mainly, djs, relation, niss, expanded, counterpart oxford, unsett, someday, mislead, hemisphere, workforce, intim, bey, naomi, wats

A.5 Occupation: Creator vs. Performer

Occupation: Creator vs. Performer	
Word	nearest neighbors in each corpus
echo	raided, adequate, pienza, provocative, bitcoin, mast, bouncing, banana, sanct, ily awar, weasel, husbands, meets, involving, somers, aussie, showcasing, duets, nutshell
inc	dwayne, higgins, recon, inclusive, exhib, bik, que, guestmix, frequency, tri bejj, inherit, li, muffins, astros, presley, semester, exercise, database, hydrated
cont	personalized, rockies, flavored, mons, santee, lovin, gofundme, monaco, islamists, tex kennedy, morris, admire, summit, verify, coup, degrees, cuis, trucks, frogs
presents	thesis, gron, backlash, bangs, mainly, djs, relation, niss, expanded, counterpart rum, disingenuous, believe, traum, submitted, ars, copying, dramas, prepar, ballin
rebel	literal, replay, fireplace, reform, bloomberg, imagination, cheer, opec, thiel, slow cartel, timmy, bahamas, function, kennedy, carol, castmates, spawn, roadtrip, bother
buck	disrupted, ruff, collections, jumps, himself, rediscover, piper, broad, lesb, jobs rede, inaugural, harlem, muster, unlimited, marshall, wearing, sterling, freaks, mindful
thee	rosie, rounded, ast, agree, wonderful, alaska, carol, atlet, vodka, neighborhoods eiffel, gutter, addict, thrones, orang, stra, slept, wiz, transit, irr
chapter	diffic, alright, smo, searched, writer, noise, raised, teamed, projecting, bud infectious, jenny, effectively, turned, deduct, sponge, worked, bankruptcy, tote, verse
dash	outweigh, edgar, wally, brows, melleefresh, repeatedly, mons, justifies, cun, rescues thankful, wearing, january, poc, billboard, dedication, clap, abi, tots, pint
op	peeps, galley, precise, persuasive, hmmm, silhouette, tree, relig, keen, communist guardian, colored, soldier, locker, boba, adorbs, pearl, dedication, demanded, le

A.6 Ideal Nearest Neighbors Results

BERT Base Uncased Nearest Neighbors	
Word	nearest neighbors in each corpus
basketball	volleyball, hockey, baseball, drama, softball, swimming, handball, sports, badminton, futsal
track	tracks, listing, song, sleeve, sleeves, synth, accompaniment, fx, peel, band
wing	squadron, plane, soaring, insect, airplane, migratory, detached, abdomen, inflated, planes