BERT Has Uncommon Sense: Similarity Ranking for Word Sense BERTology

Luke Gessler, Nathan Schneider

{lg876,nathan.schneider}@georgetown.edu

GEORGETOWN UNIVERSITY

nert.georgetown.edu

Introduction

- BERTology has surveyed the linguistic abilities of BERT and other CWE models (Rogers et al. 2020)
- Still unknown: **how well does BERT discern word senses** (especially **rare** ones)?
- We develop a **query-by-example evaluation**: given a word in context, try to find the most similar instances from a corpus

Related work

- "Word sense": a label applied to a word classifying it according to its syntax and semantics; from WSD (Navigli 2009)
- Wiedemann et al. (2019) and Reif et al. (2019) use kNN classifier with CWE models' embeddings as a WSD system
- Tayyar Madabushi et al. (2020) and Levine et al. (2020) modify BERT's training scheme with sense-oriented tasks

CWE Similarity Ranking

- Use a CWE model to embed a target token in its sentence
- Do the same for a corpus, rank corpus sentences by cosine similarity between the query token and all corpus tokens with the same lemma
- Evaluate ranking with precision at *k*
- Very similar to kNN classification, but similarity ranking awards "partial credit" – useful for rare senses where a correct kNN classification would only rarely occur

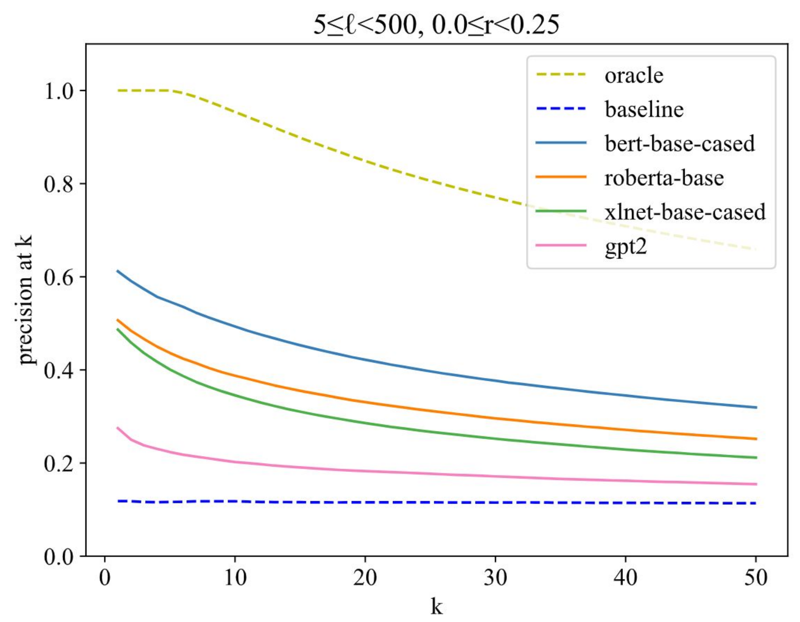
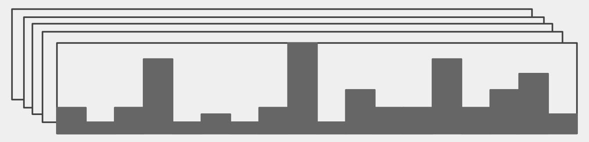


Figure 1: A sample of averaged precision at k curves, showing performance on the $\ell < 500, r < 0.25$ bucket of OntoNotes.

Query
But with all the money and glamour of high finance come the relentless pressures to do well; pressure to **pull$_4$** off another million before lunch

Target embedding

Database of embeddings



Cosine Similarity Ranking

$$\frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

Ranking

1. "Sometimes," he says, "we'll **pull$_3$** someone off phones for more training."
2. Hence, they have never lacked their own stately or amusing charms to **pull$_2$** in wealth and keep it within a household..
3. I can't **pull$_4$** it off.
4. Bulatovic says Kostunica was able to **pull$_4$** off the balancing act because he is not really anti-American.
5. ...

Experiments

- Two English corpora: OntoNotes 5.0 (Hovy et al. 2006) (**nouns and verbs**); and PDEP (Litkowski 2014), (**prepositions**)
- Compare CWE models with versions inoculated by fine-tuning (Richardson et al., 2020; Liu et al, 2019), i.e. **fine-tuned** on a **small (≤2500) number of instances from a similar task**: supersense tagging on STREUSLE (Schneider and Smith, 2015).
  - Gives model a **chance to "surface" deep information**
- Our metric: **average precision at *k* for the first 50 results**, bucketed based on lemma frequency ℓ and the sense's proportional rate of occurrence ("prevalence") r

Model	$\ell < 500$ $r < 0.25$	$\ell < 500$ $r \geq 0.25$	$\ell > 500$ $r < 0.25$	$\ell > 500$ $r \geq 0.25$
Baseline	11.55	62.41	9.55	76.08
Oracle	82.02	93.89	100.00	100.00
bert-base-cased	41.60	81.89	48.48	88.53
distilbert-base-cased	39.80	81.32	48.17	88.50
roberta-base	32.87	78.39	45.16	88.37
distilroberta-base	29.33	76.48	43.69	86.86
albert-base-v2	40.44	81.81	51.58	89.56
xlnet-base-cased	28.72	75.07	36.16	84.29
gpt2	18.34	69.56	33.53	82.74

(a) Performance for OntoNotes, no fine-tuning. Buckets respectively contain 6,949, 30,694, 1,649, and 11,123 query instances.

Table 1: Mean average precision performance. Performance is bucketed, as indicated in column headers: ℓ is a query instance's lemma frequency in the corpus, and r is the proportional frequency of a query instance's sense across all instances of the lemma in the corpus.

Results

- All **popular CWE models beat a random baseline**, even for rare senses
- But **considerable differences** in our evaluation **despite similar performance on GLUE** (Wang et al. 2019)
  - Surprising, given how similar e.g. RoBERTa and BERT are
  - Differences lessen but persist with inoculation
- Takeaway: **high-level evaluations** are informative, but **may not reveal domain-specific differences between CWE models**

References

- Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, pages 57–60, New York City, USA. Association for Computational Linguistics.
- Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In Proc. of ACL, pages 4656–4667, Online. Association for Computational Linguistics.
- Ken Litkowski. 2014. Pattern dictionary of English prepositions. In Proc. of ACL, pages 1274–1283, Baltimore, Maryland. Association for Computational Linguistics.
- Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019b. Inoculation by fine-tuning: A method for analyzing challenge datasets. In Proc. of NAACL-HLT, pages 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.
- Roberto Navigli. 2009. Word sense disambiguation: A survey. ACM Computing Surveys, 41(2):1–69.
- Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viégas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of BERT. In Proc. of NeurIPS, pages 8592–8600, Vancouver, BC, Canada.
- Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. Proc. of AAAI, 34:8713–8721.
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.
- Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In Proc. of NAACL-HLT, pages 1537–1547, Denver, Colorado. Association for Computational Linguistics.
- Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In Proc. of ICCL, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proc. of ICLR. OpenReview.net.

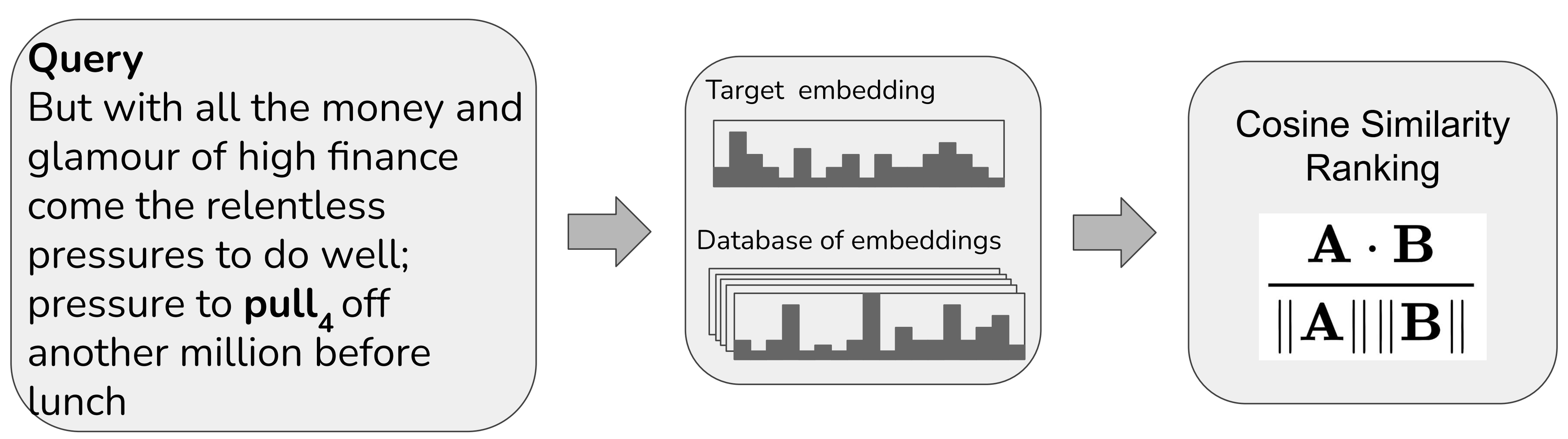
BERT Has Uncommon Sense: Similarity Ranking for Word Sense BERTology

Luke Gessler, Nathan Schneider
{lg876,nathan.schneider}@georgetown.edu

GEORGETOWN UNIVERSITY
nert.georgetown.edu

Introduction

- BERTology has surveyed the linguistic abilities of BERT and other CWE models (Rogers et al. 2020)
- Still unknown: **how well does BERT discern word senses**?
- We develop a **non-parametric similarity ranking scheme** for evaluating CWE models' word sense abilities
- We find that in English, all **popular CWE models beat a random baseline**, **even for rare senses**, but that they also **differ considerably** in their performance **despite being very similar in comprehensive benchmarks like GLUE** (Wang et al. 2019)
- Takeaway: **high-level evaluations** are informative, but **may obscure domain-specific differences between CWE models**



Results

1. "Sometimes," he says, "we'll **pull$_3$** someone off phones for more training."
2. Hence, they have never lacked their own stately or amusing charms to **pull$_2$** in wealth and keep it within a household..
3. I can't **pull$_4$** it off.
4. Bulatovic says Kostunica was able to **pull$_4$** off the balancing act because he is not really anti-American.
5. ...

CWE Similarity Ranking

- Given some CWE model *f*, an instance from a "query" corpus $\mathcal{Q}$ with some lemma *L* and instances from a "database" corpus $\mathcal{D}$ which also have the lemma *L*:
  a. Use *f* to encode the query instance's sentence, and take the sense-annotated word's embedding
  b. Use *f* to encode the database instances' sentences, and take the sense-annotated words' embeddings
  c. Use cosine similarity to rank the database instances
  d. Evaluate the ranking using precision at *k*
- Very similar to kNN classification, but similarity ranking awards "partial credit" – useful for rare senses where a correct kNN classification would only rarely occur

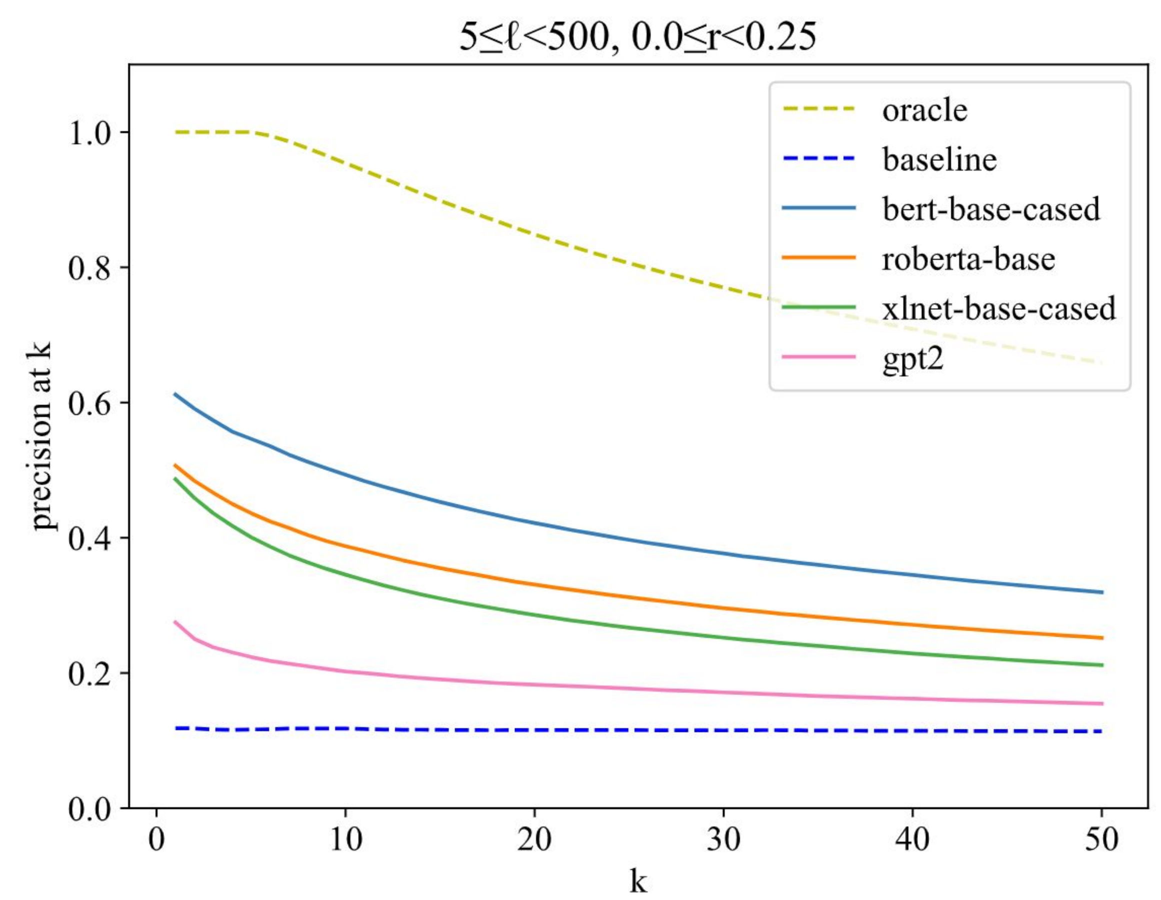


Figure 1: A sample of averaged precision at k curves, showing performance on the $\ell < 500, r < 0.25$ bucket of OntoNotes.

Experiments

- Two English corpora: OntoNotes 5.0 (Hovy et al. 2006) (**nouns and verbs**); and PDEP (Litkowski 2014), (**prepositions**)
- Compare CWE models with versions inoculated by fine-tuning (Richardson et al., 2020; Liu et al, 2019), i.e. **fine-tuned** on a **small (≤2500) number of instances from a similar task**: supersense tagging on STREUSLE (Schneider and Smith, 2015).
  - Gives model a **chance to "surface" deep information**
- Our metric: **average precision at *k* for the first 50 results**, bucketed based on lemma frequency ℓ and the sense's proportional rate of occurrence ("prevalence") *r*

Results

- Little differentiation among CWEs in the high-prevalence $r \geq 0.25$ buckets
- **Large differences for rare senses** $r < 0.25$: GPT-2 does worst, **RoBERTa remains far behind BERT even with inoculation**
- This difference is surprising, since RoBERTa is architecturally identical to BERT, differing only in training regime; even more surprising since RoBERTa slightly outperforms BERT on GLUE
- CWEs have **domain-specific performance differences** which are **not revealed by benchmarks**

Model	$\ell < 500$ $r < 0.25$	$\ell < 500$ $r \geq 0.25$	$\ell > 500$ $r < 0.25$	$\ell > 500$ $r \geq 0.25$
Baseline	11.55	62.41	9.55	76.08
Oracle	82.02	93.89	100.00	100.00
bert-base-cased	41.60	81.89	48.48	88.53
distilbert-base-cased	39.80	81.32	48.17	88.50
roberta-base	32.87	78.39	45.16	88.37
distilroberta-base	29.33	76.48	43.69	86.86
albert-base-v2	40.44	81.81	51.58	89.56
xlnet-base-cased	28.72	75.07	36.16	84.29
gpt2	18.34	69.56	33.53	82.74

(a) Performance for OntoNotes, no fine-tuning. Buckets respectively contain 6,949, 30,694, 1,649, and 11,123 query instances.

Model	$\ell < 500$ $r < 0.25$	$\ell < 500$ $r \geq 0.25$	$\ell > 500$ $r < 0.25$	$\ell > 500$ $r \geq 0.25$
Baseline	11.56	56.34	18.25	49.88
Oracle	96.41	100.00	100.00	100.00
bert-base-cased	59.59	83.54	66.34	89.39
distilbert-base-cased	58.06	83.15	65.09	88.10
roberta-base	39.84	76.79	47.65	80.01
distilroberta-base	32.42	72.22	42.29	70.81
albert-base-v2	56.53	82.22	66.64	88.44
xlnet-base-cased	35.76	74.08	37.68	75.16
gpt2	21.75	63.41	33.33	61.00

(b) Performance for PDEP, no fine-tuning. Buckets respectively contain 4,970, 1,618, 733, and 699 query instances.

Model	$\ell < 500$ $r < 0.25$	$\ell < 500$ $r \geq 0.25$	$\ell > 500$ $r < 0.25$	$\ell > 500$ $r \geq 0.25$
Baseline	11.55	62.41	9.55	76.08
Oracle	82.02	93.89	100.00	100.00
bert-base-cased	43.42	82.37	49.81	89.45
distilbert-base-cased	41.62	81.98	50.31	89.43
roberta-base	37.87	80.68	53.65	89.43
distilroberta-base	34.74	79.27	48.50	88.74
albert-base-v2	39.26	81.50	51.65	89.31
xlnet-base-cased	37.53	79.12	51.40	87.97
gpt2	18.12	68.92	32.99	82.08

(c) Best performance across all fine-tuning trials for each model on OntoNotes.

Model	$\ell < 500$ $r < 0.25$	$\ell < 500$ $r \geq 0.25$	$\ell > 500$ $r < 0.25$	$\ell > 500$ $r \geq 0.25$
Baseline	11.56	56.34	18.25	49.88
Oracle	96.41	100.00	100.00	100.00
bert-base-cased	60.37	83.49	67.87	89.28
distilbert-base-cased	58.33	82.91	67.27	87.52
roberta-base	48.99	80.32	60.04	85.72
distilroberta-base	42.25	77.25	53.37	79.19
albert-base-v2	53.75	81.85	65.57	86.97
xlnet-base-cased	49.46	80.50	56.19	84.53
gpt2	21.53	63.00	35.57	61.08

(d) Best performance across all fine-tuning trials for each model on PDEP.

Table 1: Mean average precision performance broken down by corpus and model. Performance was further measured on different buckets of instances, as indicated in column headers: ℓ is a query instance's lemma frequency in $\mathcal{D}$, and r is the proportional frequency of a query instance's sense across all instances of the lemma in $\mathcal{D}$.

Related work

- "Word sense": a label applied to a word classifying it according to its syntax and semantics; from WSD (Navigli 2009)
- Wiedemann et al. (2019) and Reif et al. (2019) use kNN classifier with CWE models' embeddings as a WSD system
- Tayyar Madabushi et al. (2020) and Levine et al. (2020) modify BERT's training scheme with sense-oriented tasks

References

- Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers, pages 57–60, New York City, USA. Association for Computational Linguistics.
- Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In Proc. of ACL, pages 4656–4667, Online. Association for Computational Linguistics.
- Ken Litkowski. 2014. Pattern dictionary of English prepositions. In Proc. of ACL, pages 1274–1283, Baltimore, Maryland. Association for Computational Linguistics.
- Nelson F. Liu, Roy Schwartz, and Noah A. Smith. 2019b. Inoculation by fine-tuning: A method for analyzing challenge datasets. In Proc. of NAACL-HLT, pages 2171–2179, Minneapolis, Minnesota. Association for Computational Linguistics.
- Roberto Navigli. 2009. Word sense disambiguation: A survey. ACM Computing Surveys, 41(2):1–69.
- Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viégas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of BERT. In Proc. of NeurIPS, pages 8592–8600, Vancouver, BC, Canada.
- Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. Proc. of AAAI, 34:8713–8721.
- Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.
- Nathan Schneider and Noah A. Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In Proc. of NAACL-HLT, pages 1537–1547, Denver, Colorado. Association for Computational Linguistics.
- Harish Tayyar Madabushi, Laurence Romain, Dagmar Divjak, and Petar Milin. 2020. CxGBERT: BERT meets construction grammar. In Proc. of ICCL, pages 4020–4032, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proc. of ICLR. OpenReview.net.