BERT Has Uncommon Sense: Similarity Ranking for Word Sense BERTology

Luke Gessler, Nathan Schneider

{lg876,nathan.schneider}@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

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 f and the sense’s proportional rate of occurrence (“prevalence”) r

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

Ranking

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

Figure 1: A sample of averaged precision at k curves, showing performance on the f ≤ 500, r ≤ 0.25 bucket of OntoNotes.

<table>
<thead>
<tr>
<th>Model</th>
<th>f ≤ 500</th>
<th>f &lt; 500</th>
<th>f ≥ 500</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>11.55</td>
<td>62.41</td>
<td>9.55</td>
</tr>
<tr>
<td>Oracle</td>
<td>82.02</td>
<td>93.89</td>
<td>100.00</td>
</tr>
<tr>
<td>bert-base</td>
<td>41.60</td>
<td>81.89</td>
<td>48.48</td>
</tr>
<tr>
<td>bert-base-cased</td>
<td>39.90</td>
<td>81.32</td>
<td>48.17</td>
</tr>
<tr>
<td>roberta-base</td>
<td>32.87</td>
<td>78.39</td>
<td>45.16</td>
</tr>
<tr>
<td>roberta-base</td>
<td>29.33</td>
<td>76.48</td>
<td>43.69</td>
</tr>
<tr>
<td>bert-base-cased</td>
<td>40.44</td>
<td>81.81</td>
<td>51.58</td>
</tr>
<tr>
<td>gpt2</td>
<td>28.72</td>
<td>75.07</td>
<td>36.16</td>
</tr>
<tr>
<td>gpt2</td>
<td>18.34</td>
<td>69.56</td>
<td>33.53</td>
</tr>
</tbody>
</table>

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

Table 1: Mean average precision performance. Performance is bucketed, as indicated in column headers: f is the proportion of a lemma’s 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.

References

- Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viégas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of words.
- Yoav Levine, Barak Lenz, Ori Dagan, Dan Padnos, Or Dagan, Ori Ram, Dan Padnos, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Evolving BERT to handle sense information.
- Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viégas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of words.
- Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viégas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of words.
- Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viégas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of words.
- Emily Reif, Ann Yuan, Martin Wattenberg, Fernanda B. Viégas, Andy Coenen, Adam Pearce, and Been Kim. 2019. Visualizing and measuring the geometry of words.
BERT Has Uncommon Sense: Similarity Ranking for Word Sense BERTology

Luke Gessler, Nathan Schneider
{lq876, nathan.schneider}@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

CWE Similarity Ranking

- Given some CWE model f, an instance from a "query" corpus Q with some lemma L and instances from a "database" corpus 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' embeddings
  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

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
- Our metric: average precision at k for the first 50 results, bucketed based on lemma frequency f, and the sense's proportional rate of occurrence ("prevalence") r

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

Results

- Little differentiation among CWEs in the high-prevalence r ≥ 0.25 buckets
- 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

References