Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information

Michele Bevilacqua and Roberto Navigli

ACL2020
Overview

• Worse Sense Disambiguation: given a word \(w\) and a context \(c\), predict the sense (synset) of the word.

• This work (EWISER) extends the idea of the EWISE (ACL2019) model in:
  + Computing the unnormalized scores of the words (logits) \(z\).
  + Using synset embeddings \(O\) in the output layer.
\[ z_s = h^T g(s) + b^T g(s) \]
EWISER: Baseline

• Use BERT to produce context-aware representations for words.
• The unnormalized scores of the words $z$ are computed by:

\[
B = B_{-4} + B_{-3} + B_{-2} + B_{-1}
\]

\[
H_0 = \text{BatchNorm}(B)
\]

\[
H_1 = \text{swish}(H_0 W + b)
\]

\[
Z = H_1 O
\]
EWISER: Structured logits

• Using Lexical Knowledge Bases (LKBs) with relational information between synsets to obtain different structures.

• The structures are encoded by an adjacency matrix $A$, which can be learned or fixed during training.

• The score of a word w.r.t a sense (synset) now also depends on the neighbors of the main synset in the structure encoded in $A$.

$$q_s = z_s + \sum_{s' \in V | (s', s) \in E} w((s', s)) \cdot z_{s'}$$
EWISER: Structured logits

- This work experiments with different relations existing in WordNet to obtain different types of A.

| Model Arch. | ALL | No15 | No15* |
|-------------|-----|------|-------|
| baseline    | 74.2| 73.9 | 52.2  |
| hyper       |     |      |       |
|             | 75.6| 75.4 | 59.8  |
|             | 75.9| 75.5 | 59.2  |
| hypo        |     |      |       |
|             | 74.6| 74.4 | 57.7  |
|             | 74.6| 74.3 | 54.5  |
| hyper+hypo  |     |      |       |
|             | 75.7| 75.5 | 59.8  |
|             | 75.7| 75.4 | 57.7  |
| hyper*      |     |      |       |
|             | 75.2| 75.0 | 58.6  |
|             | 75.4| 75.3 | 57.7  |
| hyper+hypo* |    |      |       |
|             | 75.4| 75.3 | 59.9  |
|             | 74.7| 74.4 | 56.5  |

Table 1: Evaluation of structured logits on English all-words WSD. F1 is reported.
EWISER: Synset Embeddings

- Try different strategies for incorporating the synset embeddings into the network:
  - Init: plain initialization
  - Freeze: Pretrained initialization and freeze.
  - Thaw: Training a freeze model, restore the best checkpoint, further training "thawed".
  - Thaw*: same as "Thaw", but with a smaller learning rate.
EWISER: Synset Embeddings

| Model Arch.     | ALL  | No15 | No15^- |
|-----------------|------|------|--------|
| baseline        | 74.2 | 73.9 | 52.2   |
| Deconf          |      |      |        |
| O-init          | 75.3 | 75.2 | 55.2   |
| O-freeze        | 66.4 | 66.0 | 72.2   |
| O-thaw          | 75.3 | 75.2 | 60.5   |
| O-thaw*         | 73.8 | 73.7 | 62.3   |
| LMMS            |      |      |        |
| O-init          | 75.5 | 75.4 | 55.1   |
| O-freeze        | 75.9 | 75.4 | 59.4   |
| O-thaw          | 75.4 | 75.0 | 57.4   |
| O-thaw*         | 75.8 | 75.4 | 57.3   |
| LMMS + SensEmBERT|     |      |        |
| O-init          | 76.1 | 76.0 | 59.4   |
| O-freeze        | 76.3 | 76.0 | 64.7   |
| O-thaw          | 76.4 | 76.1 | 62.3   |
| O-thaw*         | **76.7** | **76.6** | **63.4** |

Table 2: Evaluation of O initialization and training strategies on English all-words WSD. F1 is reported.
## Results

| S | G | G⁺ | E | System                                      | ALL  | No15 | No15⁻ | S2  | S3  | S7  | S13 | S15 | N   | V  | A  | R   |
|---|---|----|---|-------------------|------|------|-------|-----|-----|-----|-----|-----|-----|----|----|-----|
| ✓ | ✓ | -  | - | Kumar et al. (2019)     | 71.8 | 70.9*| -     | 73.8| 71.1| 67.3| 69.4| 74.5| 74.0| 60.2| 78.0| 82.1|
| ✓ | ✓ | -  | - | Loureiro and Jorge (2019) | 75.4 | 75.2*| -     | 76.3| 75.6| 68.1| 75.1| 77.0| -   | -   | -   | -   |
| ✓ | - | -  | - | Hadiwinoto et al. (2019)  | 73.7*| 73.2*| -     | 75.5| 73.6| 68.1| 71.1| 76.2| -   | -   | -   | -   |
| ✓ | ✓ | -  | - | Huang et al. (2019)       | 77.0*| 76.2*| -     | 77.7| 75.2| 72.5| 76.1| 80.4| -   | -   | -   | -   |
| ✓ | ✓ | -  | - | Scarlini et al. (2020) - Sup. | -   | -   | -     | -   | -   | -   | -   | 78.7 | 80.4 | -   | -   | -   |
| ✓ | - | -  | - | Vial et al. (2019)        | 75.6 | -    | -     | -   | -   | -   | -   | -   | -   | -   | -   | -   |
| ✓ | - | -  | - | Vial et al. (2019) - ENS  | 76.7 | 76.5*| -     | 77.5| 77.4| 69.5| 76.0| 78.3| 79.6| 65.9| 79.5| 85.5|
| ✓ | † | -  | - | EWISER_{hyper}            | 77.0*| 76.9 | 60.4  | 77.5| 77.9| 71.0| 76.4| 77.8| 79.9| 66.4| 79.0| 85.5|
| ✓ | - | -  | - | EWISER_{hyper}            | 77.5 | 77.3 | 68.2  | 78.4| 77.4| 71.0| 77.4| 78.7| 80.7| 65.1| 80.9| 86.1|
| ✓ | † | -  | - | EWISER_{hyper+hypo}      | 76.8 | 76.8 | 59.5  | 77.7| 77.9| 70.3| 76.2| 76.3| 79.4| 65.9| 80.0| 86.7|
| ✓ | ✓ | -  | - | EWISER_{hyper+hypo}      | 78.3 | 78.2 | 69.1  | 78.9| 78.4| 71.0| 78.9| 79.3| 81.7| 66.3| 81.2| 85.8|
| ✓ | ✓ | ✓ | ✓ | Vial et al. (2019)        | 77.1 | -    | -     | -   | -   | -   | -   | -   | 81.7| 66.3| 81.2| 85.8| -   |
| ✓ | ✓ | ✓ | ✓ | Vial et al. (2019) - ENS  | 79.0*| 78.4*| -     | 79.7| 77.8| 73.4| 78.7| 82.6| 81.4| 68.7| 83.7| 85.5| -   |
| ✓ | ✓ | ✓ | ✓ | EWISER_{hyper}            | 80.1 | 79.8 | 75.2  | 80.8| 79.0| 75.2| 80.7| 81.8| 82.9| 69.4| 83.6| 87.3| -   |
| ✓ | ✓ | ✓ | ✓ | EWISER_{hyper+hypo}      | 79.8 | 79.3 | 75.1  | 80.2| 78.5| 73.8| 80.6| 82.3| 82.7| 68.5| 82.9| 87.6| -   |
| - | - | -  | - | Scozzafava et al. (2020)  | 71.7 | 71.0*| -     | 71.6| 72.0| 59.3| 72.2| 75.8| -   | -   | -   | -   |
| - | ✓ | -  | - | Scarlini et al. (2020) - KB| -   | -    | -     | -   | 74.8| -   | -   | -   | 75.9 | -   | -   | -   | -   |
## Results

- Cross-lingual WSD:

|                      | S13 DE | S13 ES | S13 FR | S13 IT | S15 ES | S15 IT |
|----------------------|--------|--------|--------|--------|--------|--------|
| Scozzafava et al. (2020) | 76.4   | 74.1   | 70.3   | 72.1   | 63.4   | 69.0   |
| Scarlini et al. (2020)    | 79.2*  | 73.4*  | 77.8*  | 69.8*  | -      | -      |
| Ours (baseline)         | 81.7   | 76.6   | 80.8   | 77.2   | 67.3   | 70.6   |
| Ours (EWISER)          | 80.9   | **78.8** | **83.6** | **77.7** | 69.5   | **71.8** |