Breaking Through the 80% Glass Ceiling:

Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information

Michele Bevilacqua   Roberto Navigli

bevilacqua,navigli@di.uniroma1.it
Sapienza NLP
Department of Computer Science
Sapienza University of Rome

ERC Consolidator Grant MOUSSE No. 726487 under the European Union's Horizon 2020 research and innovation programme. This work was supported in part by the MIUR under the grant "Dipartimenti di eccellenza 2018-2022" of the Department of Computer Science of the Sapienza University of Rome.

What is Word Sense Disambiguation?

- Words are ambiguous:

  A **bat** is flying towards you at full speed!

What is Word Sense Disambiguation?

- Words are ambiguous:

  A **bat** is flying towards you at full speed!

- Word Sense Disambiguation (WSD) frames **polysemy resolution as a multi-class classification problem**.

What is Word Sense Disambiguation?

- Words are ambiguous:

  A **bat** is flying towards you at full speed!

- Word Sense Disambiguation (WSD) frames **polysemy resolution as a multi-class classification problem**.
- Classes (WordNet 3.0):

SAPIENZA NLP

What is Word Sense Disambiguation?

- Words are ambiguous:

  A **bat** is flying towards you at full speed!

- Word Sense Disambiguation (WSD) frames **polysemy resolution as a multi-class classification problem**.
- Classes (WordNet 3.0):
  - *nocturnal mouselike mammal with forelimbs modified to form membranous wings [...]*

SAPIENZA NLP

What is Word Sense Disambiguation?

- Words are ambiguous:

  A **bat** is flying towards you at full speed!

- Word Sense Disambiguation (WSD) frames **polysemy resolution as a multi-class classification problem**.
- Classes (WordNet 3.0):
  - *nocturnal mouselike mammal with forelimbs modified to form membranous wings [. . .]*
  - *(baseball) a turn trying to get a hit*

SAPIENZA NLP

What is Word Sense Disambiguation?

- Words are ambiguous:

  A **bat** is flying towards you at full speed!

- Word Sense Disambiguation (WSD) frames **polysemy resolution as a multi-class classification problem**.
- Classes (WordNet 3.0):
  - *nocturnal mouselike mammal with forelimbs modified to form membranous wings [...]*
  - *(baseball) a turn trying to get a hit*
  - *a club used for hitting a ball in various games*



SAPIENZA NLP

Supervised WSD Systems

Traditional supervised WSD approaches [e.g., Hadiwinoto et al., 2019] treat senses as **opaque classes** and learning to associate words in context with senses using **only training data**.

Supervised WSD Systems

Traditional supervised WSD approaches [e.g., Hadiwinoto et al., 2019] treat senses as **opaque classes** and learning to associate words in context with senses using **only training data**.

Coverage: senses often missing in training data! $\rightarrow$ back-off strategies.

Supervised WSD Systems

Traditional supervised WSD approaches [e.g., Hadiwinoto et al., 2019] treat senses as **opaque classes** and learning to associate words in context with senses using **only training data**.

Coverage: senses often missing in training data! $\rightarrow$ back-off strategies.

Knowledge: cannot exploit the knowledge that is present in Lexical Knowledge Bases (e.g. WordNet):

SAPIENZA NLP

Supervised WSD Systems

Traditional supervised WSD approaches [e.g., Hadiwinoto et al., 2019] treat senses as **opaque classes** and learning to associate words in context with senses using **only training data**.

Coverage: senses often missing in training data! $\rightarrow$ back-off strategies.

Knowledge: cannot exploit the knowledge that is present in Lexical Knowledge Bases (e.g. WordNet):

- unstructured information (e.g. definitions);

SAPIENZA NLP

Supervised WSD Systems

Traditional supervised WSD approaches [e.g., Hadiwinoto et al., 2019] treat senses as **opaque classes** and learning to associate words in context with senses using **only training data**.

Coverage: senses often missing in training data! → back-off strategies.

Knowledge: cannot exploit the knowledge that is present in Lexical Knowledge Bases (e.g. WordNet):

- unstructured information (e.g. definitions);
- structured information (relations between senses or *synsets*).

SAPIENZA NLP

Unstructured Knowledge in WSD

- Synsets in WordNet are provided with **glosses/definitions**.

Unstructured Knowledge in WSD

- Synsets in WordNet are provided with **glosses/definitions**.
- Many previous supervised WSD approaches needed **specialized architectures** to exploit glosses:

Unstructured Knowledge in WSD

- Synsets in WordNet are provided with **glosses/definitions**.
- Many previous supervised WSD approaches needed **specialized architectures** to exploit glosses:
  - Huang et al. [2019, GlossBERT]: predict whether a gloss fits a word in context;

Unstructured Knowledge in WSD

- Synsets in WordNet are provided with **glosses/definitions**.
- Many previous supervised WSD approaches needed **specialized architectures** to exploit glosses:
    - Huang et al. [2019, GlossBERT]: predict whether a gloss fits a word in context;
    - Kumar et al. [2019, EWISE]: maps a target in context vector to the space of gloss embeddings.

Structured Knowledge

WordNet is not just a sense list, but a **directed graph**. Edges represent lexical (semantic) relations between senses (synsets).

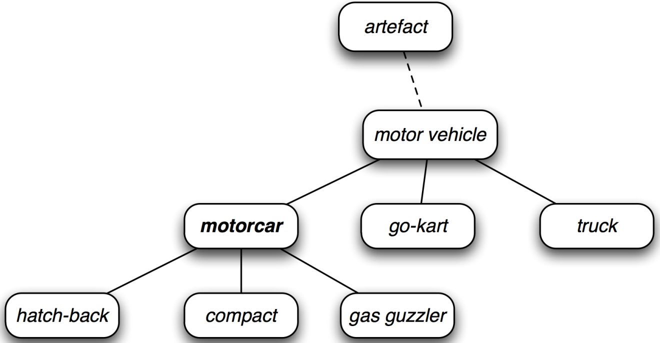


Figure: Image taken from `https://www.nltk.org/book/ch02.html`

SAPIENZA NLP

Structured Knowledge in WSD

- Graph relations are **commonly exploited by knowledge-based approaches to WSD**:
  - e.g. Personalized PageRank approaches [Agirre and Soroa, 2009].

Structured Knowledge in WSD

- Graph relations are **commonly exploited by knowledge-based approaches to WSD**:
  - e.g. Personalized PageRank approaches [Agirre and Soroa, 2009].
- **Structured knowledge is not commonplace in supervised WSD**. Few approaches have made use of these relations:

SAPIENZA NLP

Structured Knowledge in WSD

- Graph relations are **commonly exploited by knowledge-based approaches to WSD**:
  - e.g. Personalized PageRank approaches [Agirre and Soroa, 2009].
- **Structured knowledge is not commonplace in supervised WSD**. Few approaches have made use of these relations:
  - Kumar et al. [2019, EWISE]: uses relations in the triplet loss that is used to train the definition encoder. The relation information is only stored implicitly in the parameters.

SAPIENZA NLP

Structured Knowledge in WSD

- Graph relations are **commonly exploited by knowledge-based approaches to WSD**:
  - e.g. Personalized PageRank approaches [Agirre and Soroa, 2009].
- **Structured knowledge is not commonplace in supervised WSD**. Few approaches have made use of these relations:
  - Kumar et al. [2019, EWISE]: uses relations in the triplet loss that is used to train the definition encoder. The relation information is only stored implicitly in the parameters.
  - Vial et al. [2019]: relations are used to conflate senses into coarser but reversible semantic classes. Information which is specific to a synset is lost.

SAPIENZA NLP

EWISER: High-level View

- In our approach, EWISER (*Enhanced WSD Integrating Synset Embeddings and Relations*) we exploit **semantic knowledge both implicitly and explicitly**:

EWISER: High-level View

- In our approach, EWISER (*Enhanced WSD Integrating Synset Embeddings and Relations*) we exploit **semantic knowledge both implicitly and explicitly**:
    - implicit knowledge, through the use of synset embeddings;

SAPIENZA NLP

EWISER: High-level View

- In our approach, EWISER (*Enhanced WSD Integrating Synset Embeddings and Relations*) we exploit **semantic knowledge both implicitly and explicitly**:
  - implicit knowledge, through the use of synset embeddings;
  - explicit knowledge, through the incorporation of a WordNet-based adjacency matrix.

EWISER: High-level View

- In our approach, EWISER (*Enhanced WSD Integrating Synset Embeddings and Relations*) we exploit **semantic knowledge both implicitly and explicitly**:
  - implicit knowledge, through the use of synset embeddings;
  - explicit knowledge, through the incorporation of a WordNet-based adjacency matrix.
- Both techniques are added **on top of a baseline neural classifier**.

SAPIENZA NLP

EWISER: Baseline Classifier

- We use a very simple **2-layer feedforward WSD classifier**, taking as input BERT large's hidden states for of the last 4 layers:

SAPIENZA NLP

EWISER: Baseline Classifier

- We use a very simple **2-layer feedforward WSD classifier**, taking as input BERT large's hidden states for of the last 4 layers:

$$
\begin{aligned}
B &= B_{-4} + B_{-3} + B_{-2} + B_{-1} \\
H_0 &= \text{BatchNorm}(B) \\
H_1 &= \text{swish}(H_0 W + \vec{b}) \\
Z &= H_1 \cdot O
\end{aligned}
$$

SAPIENZA NLP

- We use a very simple **2-layer feedforward WSD classifier**, taking as input BERT large's hidden states for of the last 4 layers:

$$
\begin{aligned}
B &= B_{-4} + B_{-3} + B_{-2} + B_{-1} \\
H_0 &= \mathrm{BatchNorm}(B) \\
H_1 &= \mathrm{swish}(H_0 W + \vec{b}) \\
Z &= H_1 \cdot \boxed{O}
\end{aligned}
$$

Implicit Knowledge $\leftarrow$ **Initialize** $\boxed{\mathbf{O}}$ with sense embeddings.

EWISER: Baseline Classifier

- We use a very simple **2-layer feedforward WSD classifier**, taking as input BERT large's hidden states for of the last 4 layers:

$$
\begin{aligned}
B &= B_{-4} + B_{-3} + B_{-2} + B_{-1} \\
H_0 &= \mathrm{BatchNorm}(B) \\
H_1 &= \mathrm{swish}(H_0 W + \vec{b}) \\
Z &= H_1 \cdot \boxed{O} \\
Q &= \boxed{ZA^T} + Z
\end{aligned}
$$

Implicit Knowledge ← **Initialize** $\boxed{\mathbf{O}}$ with sense embeddings.

Explicit Knowledge ← Add an additional term to Z **computed via the adjacency matrix** A.

SAPIENZA NLP

EWISER: Unstructured Knowledge

- O is a h by $|V|$ matrix:
  - h: hidden size;
  - V: set of all 117659 synsets in WordNet.

SAPIENZA NLP

- O is a h by $|V|$ matrix:
  - h: hidden size;
  - V: set of all 117659 synsets in WordNet.
  - Each column vector in O is a synset $\rightarrow$ **the logit for a synset is the scalar product between the hidden vector and the synset vector**.

SAPIENZA NLP

EWISER: Unstructured Knowledge

- O is a h by $|V|$ matrix:
  - h: hidden size;
  - V: set of all 117659 synsets in WordNet.
  - Each column vector in O is a synset $\rightarrow$ **the logit for a synset is the scalar product between the hidden vector and the synset vector**.
- synset embeddings are used to provide a **better initialization than random for** O:
  - embeddings are reduced to 512 dimensions by SVD;
  - synset embeddings $\rightarrow$ centroid of sense embeddings.

SAPIENZA NLP

- O is a h by $|V|$ matrix:
  - h: hidden size;
  - V: set of all 117659 synsets in WordNet.
  - Each column vector in O is a synset $\rightarrow$ **the logit for a synset is the scalar product between the hidden vector and the synset vector**.
- synset embeddings are used to provide a **better initialization than random for** O:
  - embeddings are reduced to 512 dimensions by SVD;
  - synset embeddings $\rightarrow$ centroid of sense embeddings.
- we employ sense embeddings that **incorporate gloss information**:
  - LMMS_{2048} [Loureiro and Jorge, 2019];
  - SensEmBERT [Scarlini et al., 2020] + LMMS_{2048}.

SAPIENZA NLP

EWISER: Unstructured Knowledge - Weight Updates

- Should O **be kept fixed or updated**?

EWISER: Unstructured Knowledge - Weight Updates

- Should O **be kept fixed or updated**?
  - If we freeze, the embeddings might be suboptimal for WSD.

SAPIENZA NLP

EWISER: Unstructured Knowledge - Weight Updates

- Should O **be kept fixed or updated**?
  - If we freeze, the embeddings might be suboptimal for WSD.
  - If we train, knowledge in the embeddings might be lost.

EWISER: Unstructured Knowledge - Weight Updates

- Should O **be kept fixed or updated**?
  - If we freeze, the embeddings might be suboptimal for WSD.
  - If we train, knowledge in the embeddings might be lost.
- We evaluate different strategies to tackle this **trade-off**:

Baseline	O randomly initialized

SAPIENZA NLP

EWISER: Unstructured Knowledge - Weight Updates

- Should O **be kept fixed or updated**?
  - If we freeze, the embeddings might be suboptimal for WSD.
  - If we train, knowledge in the embeddings might be lost.
- We evaluate different strategies to tackle this **trade-off**:

Baseline	O randomly initialized
O-init	O initialized

SAPIENZA NLP

EWISER: Unstructured Knowledge - Weight Updates

- Should O **be kept fixed or updated**?
  - If we freeze, the embeddings might be suboptimal for WSD.
  - If we train, knowledge in the embeddings might be lost.
- We evaluate different strategies to tackle this **trade-off**:

Baseline	O randomly initialized
O-init	O initialized
O-freeze	O initialized and freezed

SAPIENZA NLP

EWISER: Unstructured Knowledge - Weight Updates

- Should O **be kept fixed or updated**?
  - If we freeze, the embeddings might be suboptimal for WSD.
  - If we train, knowledge in the embeddings might be lost.
- We evaluate different strategies to tackle this **trade-off**:

Baseline	O randomly initialized
O-init	O initialized
O-freeze	O initialized and freezed
O-thaw	O-freeze, then weights unfreezed

SAPIENZA NLP

EWISER: Unstructured Knowledge - Weight Updates

- Should O **be kept fixed or updated**?
  - If we freeze, the embeddings might be suboptimal for WSD.
  - If we train, knowledge in the embeddings might be lost.
- We evaluate different strategies to tackle this **trade-off**:

Baseline	O randomly initialized
O-init	O initialized
O-freeze	O initialized and freezed
O-thaw	O-freeze, then weights unfreezed
O-thaw*	O-thaw with LR reduced

SAPIENZA NLP

EWISER: Experimental Setup

- **Training**:
  - **SemCor** [Miller et al., 1994];
  - WordNet **Tagged Glosses** + WordNet Examples.

SAPIENZA NLP

EWISER: Experimental Setup

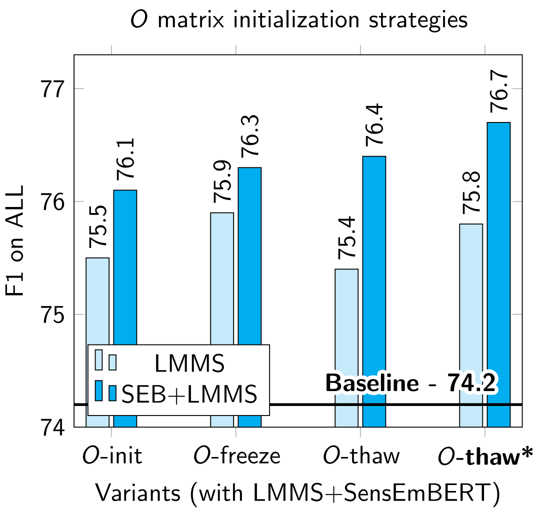
- **Training**:
  - **SemCor** [Miller et al., 1994];
  - WordNet **Tagged Glosses** + WordNet Examples.
- **Development**:
  - SemEval 2015 Task 13 [Moro and Navigli, 2015].

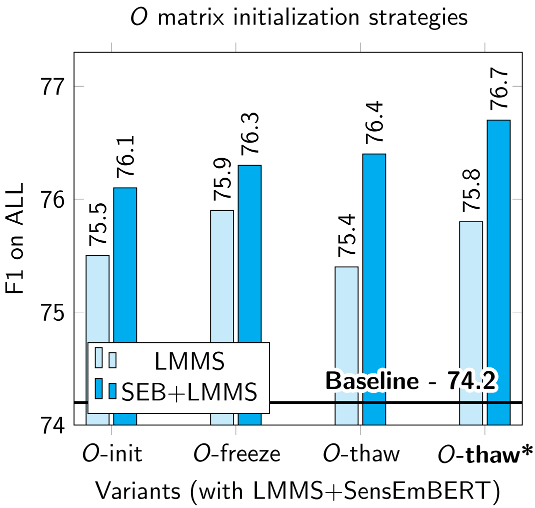
EWISER: Experimental Setup

- **Training**:
  - **SemCor** [Miller et al., 1994];
  - WordNet **Tagged Glosses** + WordNet Examples.
- **Development**:
  - SemEval 2015 Task 13 [Moro and Navigli, 2015].
- **Test**:
  - **Concatenation of the English datasets** in the framework of [Raganato et al., 2017, **ALL**];
  - **Multilingual datasets** from SemEval 2013 Task 12 [Navigli et al., 2013] and SemEval 2015 Task 13;
  - Results on individual datasets reported in the paper!

SAPIENZA NLP

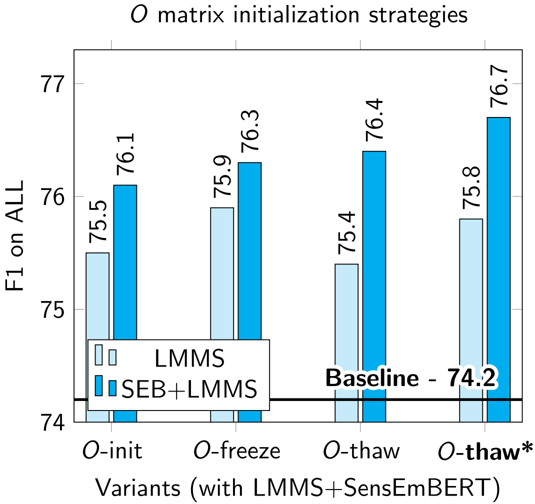
EWISER: Unstructured Knowledge - Results





- The **choice of the embeddings is critical** (SEB > LMMS);

EWISER: Unstructured Knowledge - Results



- The **choice of the embeddings is critical** (SEB > LMMS);
- ***O*-thaw*** is a very effective strategy.

EWISER: Structured Knowledge

- Structured knowledge added by a **matrix multiplication** between:
  - Z: logits;
  - A: sparse adjacency matrix.

$$Z = H_1 O$$

$$Q = Z + \boxed{ZA^T}$$

SAPIENZA NLP

EWISER: Structured Knowledge

- Structured knowledge added by a **matrix multiplication** between:
  - Z: logits;
  - A: sparse adjacency matrix.

$$Z = H_1 O$$

$$Q = Z + \boxed{ZA^T}$$

- This is equivalent to adding to a synset logit the sum of all the scores for all synsets connected to it, weighted by the edge weight:

$$\vec{q}_s = \vec{z}_s + \boxed{\sum_{s' \in V | \langle s', s \rangle \in E} w(\langle s', s \rangle) \cdot \vec{z}_{s'}}$$

SAPIENZA NLP

EWISER: Structured Knowledge

- Structured knowledge added by a **matrix multiplication** between:
  - Z: logits;
  - A: sparse adjacency matrix.

$$Z = H_1 O$$

$$Q = Z + \boxed{ZA^T}$$

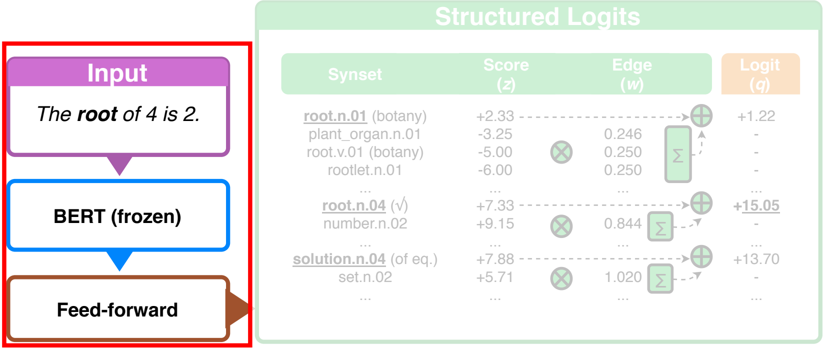
- This is equivalent to adding to a synset logit the sum of all the scores for all synsets connected to it, weighted by the edge weight:

$$\vec{q}_s = \vec{z}_s + \boxed{\sum_{s' \in V | \langle s', s \rangle \in E} w(\langle s', s \rangle) \cdot \vec{z}_{s'}}$$

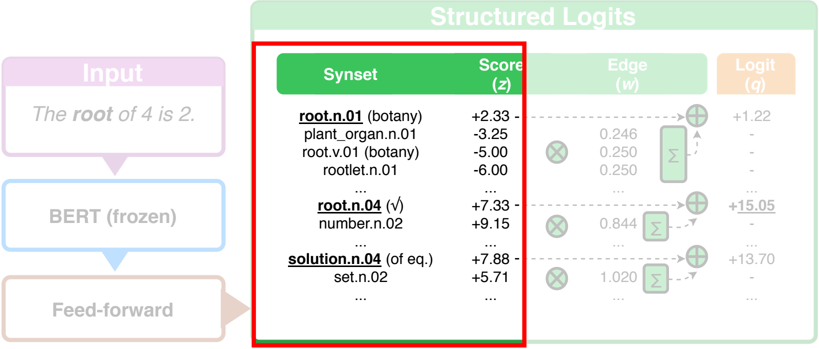
- The adjacency matrix A weights can be **refined with standard backpropagation**!

SAPIENZA NLP

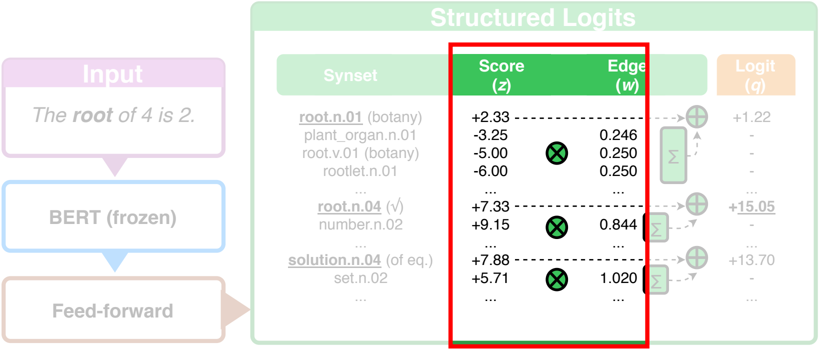
EWISER: Structured Knowledge



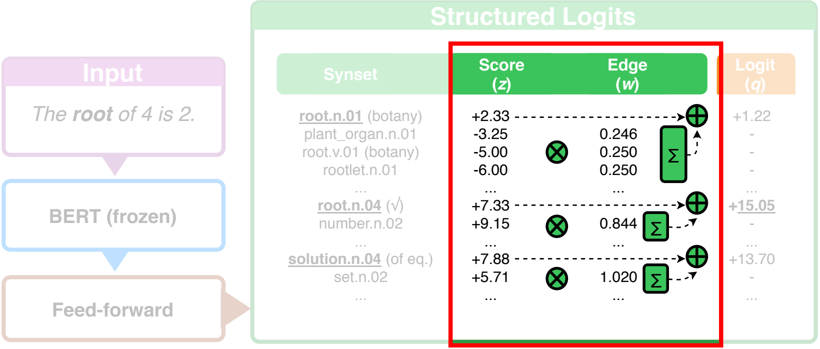
EWISER: Structured Knowledge



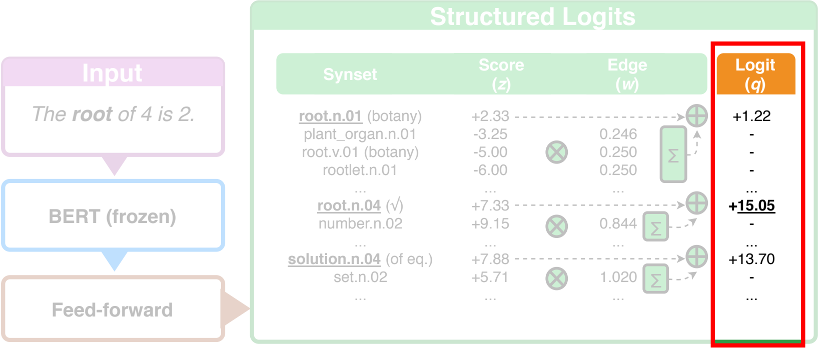
EWISER: Structured Knowledge



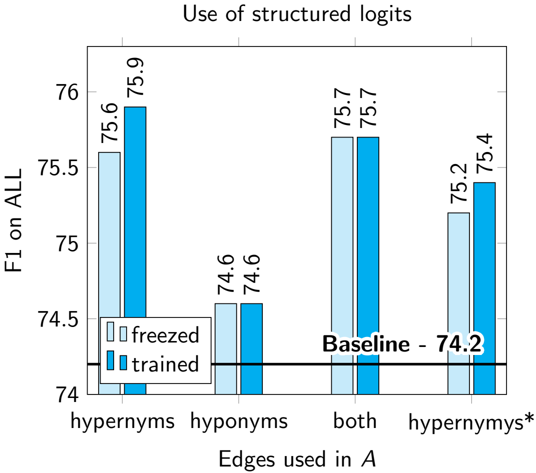
EWISER: Structured Knowledge



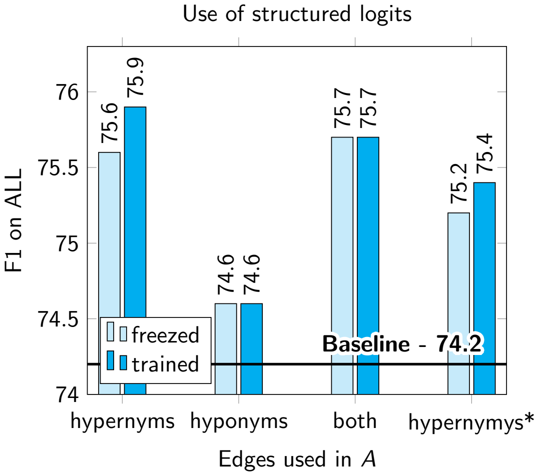
EWISER: Structured Knowledge



EWISER: Structured Knowledge - Results

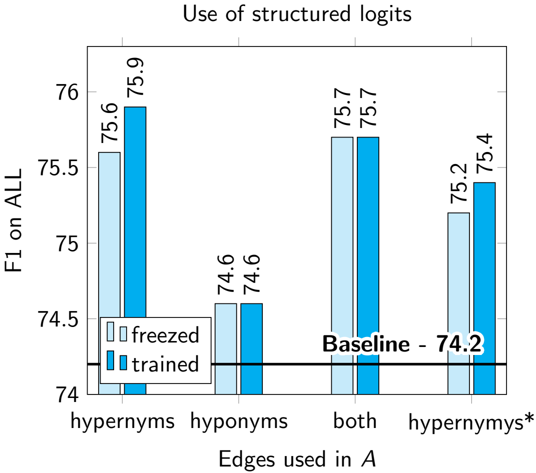


EWISER: Structured Knowledge - Results



- **hypernymy edges** must be used;

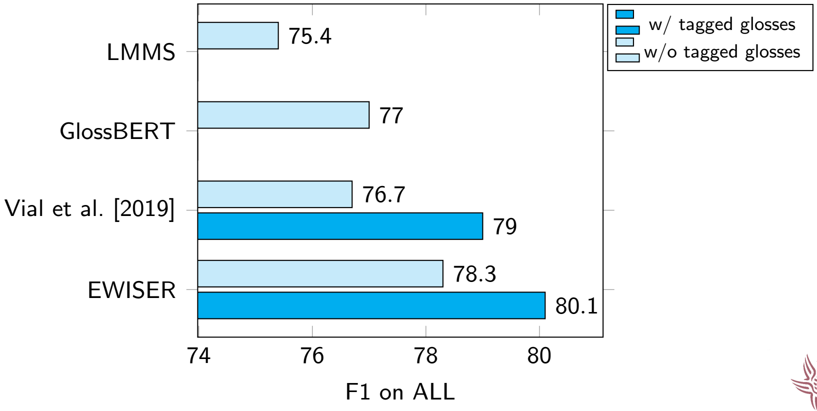
EWISER: Structured Knowledge - Results



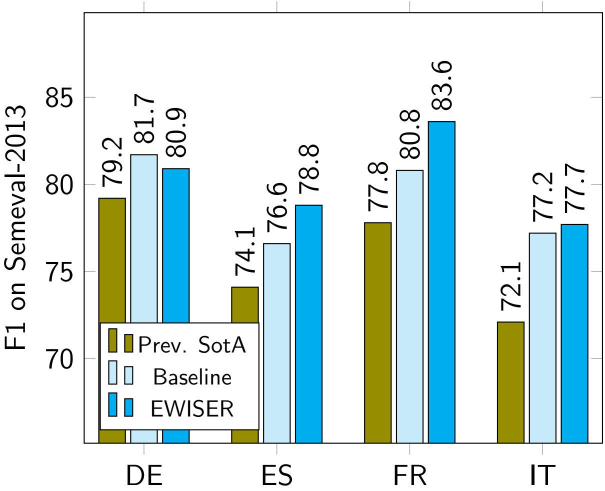
- **hypernymy edges** must be used;
- **training the edge weights** results in a small improvement.

EWISER: Bringing Everything Together

The improvements can be stacked, with **SotA results** on the concatenation of the standard evaluation datasets!



The results are also **strong in a cross-lingual setting**, with the model **trained only on English**.



Conclusion

We brought together structured and unstructured knowledge in a single WSD architecture:

Conclusion

We brought together structured and unstructured knowledge in a single WSD architecture:

- sense embeddings used to **initialize the output embeddings**;

SAPIENZA NLP

Conclusion

We brought together structured and unstructured knowledge in a single WSD architecture:

- sense embeddings used to **initialize the output embeddings**;
- relational knowledge integrated through via an **explicit WordNet adjacency matrix**;

SAPIENZA NLP

Conclusion

We brought together structured and unstructured knowledge in a single WSD architecture:

- sense embeddings used to **initialize the output embeddings**;
- relational knowledge integrated through via an **explicit WordNet adjacency matrix**;
- we **surpass for the first time the 80% figure** (upper bound on human annotator agreement on WSD) on the standard English benchmarks;

Conclusion

We brought together structured and unstructured knowledge in a single WSD architecture:

- sense embeddings used to **initialize the output embeddings**;
- relational knowledge integrated through via an **explicit WordNet adjacency matrix**;
- we **surpass for the first time the 80% figure** (upper bound on human annotator agreement on WSD) on the standard English benchmarks;
- performances **scale gracefully to the multilingual** setting.

SAPIENZA NLP

Thank you!

ERC Consolidator Grant MOUSSE No. 726487.

SAPIENZA NLP

Eneko Agirre and Aitor Soroa. Personalizing PageRank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 33–41, Athens, Greece, March 2009. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E09-1005.

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. Improved word sense disambiguation using pre-trained contextualized word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5297–5306, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1533. URL https://www.aclweb.org/anthology/D19-1533.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. GlossBERT: BERT for word sense disambiguation with gloss knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3500–3505, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1355. URL https://www.aclweb.org/anthology/D19-1355.

References II

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1568. URL `https://www.aclweb.org/anthology/P19-1568`.

Daniel Loureiro and Alípio Jorge. Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1569. URL `https://www.aclweb.org/anthology/P19-1569`.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. Using a semantic concordance for sense identification. In *Proceedings of HUMAN LANGUAGE TECHNOLOGY: a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994. URL `https://www.aclweb.org/anthology/H94-1046`.

SAPIENZA NLP

Andrea Moro and Roberto Navigli. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado, June 2015. Association for Computational Linguistics. doi: 10.18653/v1/S15-2049. URL https://www.aclweb.org/anthology/S15-2049.

Roberto Navigli, David Jurgens, and Daniele Vannella. SemEval-2013 task 12: Multilingual word sense disambiguation. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/S13-2040.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E17-1010.

SAPIENZA NLP

References IV

Bianca Scarlini, Tommaso Pasini, and Roberto Navigli. SensEmBERT: Context-enhanced sense embeddings for multilingual word sense disambiguation. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020*, 2020. URL http://sensembert.org/resources/scarlini_etal_aaai2020.pdf.

Loïc Vial, Benjamin Lecouteux, and Didier Schwab. Sense vocabulary compression through the semantic knowledge of WordNet for neural word sense disambiguation. In *Proceedings of the Global WordNet Conference*, pages 108–117, 2019. URL https://arxiv.org/abs/1905.05677.