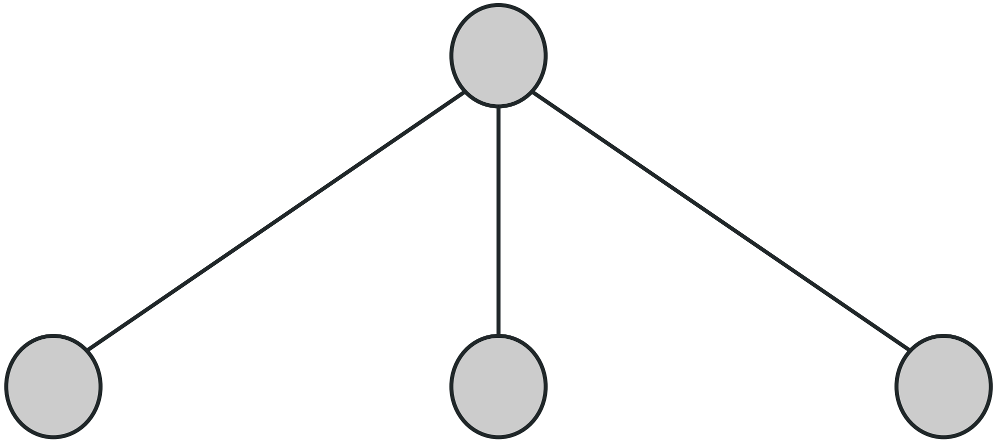
Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders

Terra Blevins and Luke Zettlemoyer

UWNLP

facebook
AI Research

The **plant** sprouted a new leaf.

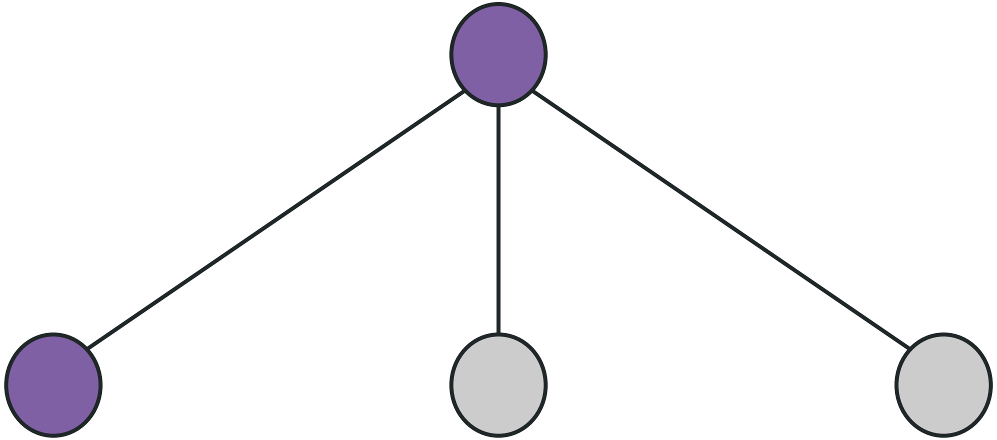


(n) (botany) a living organism…

(n) buildings for carrying on industrial labor

(v) to put or set (a seed or plant) into the ground

The **plant** sprouted a new leaf.



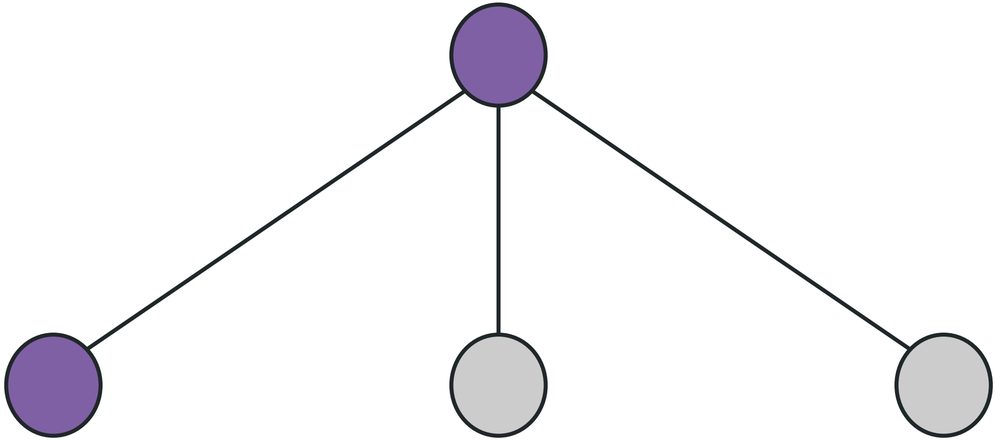
(n) (botany) a living organism...

(n) buildings for carrying on industrial labor

(v) to put or set (a seed or plant) into the ground

Target Word

Context The **plant** sprouted a new leaf.



(n) (botany) a living organism...

(n) buildings for carrying on industrial labor

(v) to put or set (a seed or plant) into the ground

Candidate Senses

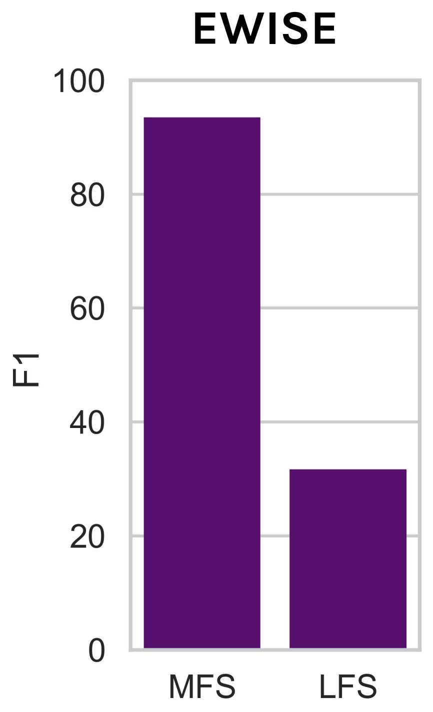
Data Sparsity in WSD

- Senses have Zipfian distribution in natural language

Kilgarriff (2004), *How dominant is the commonest sense of a word?*.
Kumar et al. (2019), *Zero-shot Word Sense Disambiguation using Sense Definition Embeddings.*

Data Sparsity in WSD

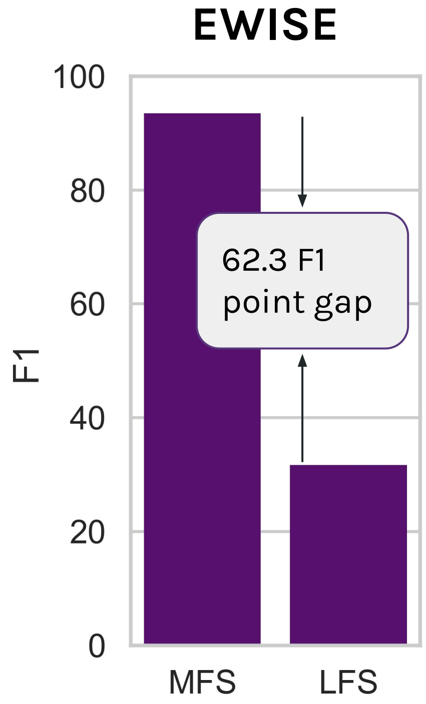
- Senses have Zipfian distribution in natural language
- Data imbalance leads to worse performance on uncommon senses



Kilgarriff (2004), *How dominant is the commonest sense of a word?.*
Kumar et al. (2019), *Zero-shot Word Sense Disambiguation using Sense Definition Embeddings.*

Data Sparsity in WSD

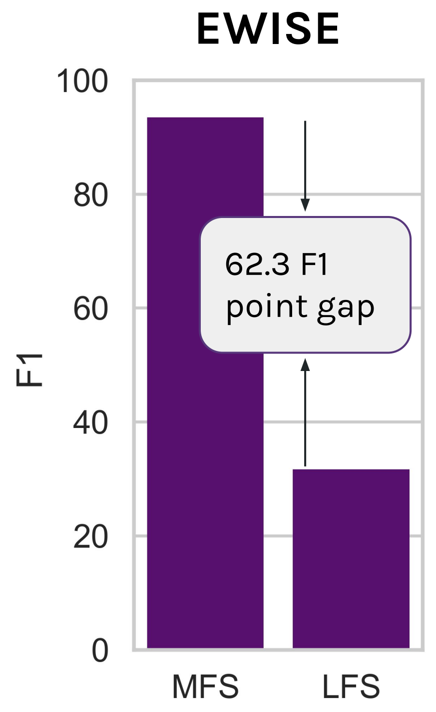
- Senses have Zipfian distribution in natural language
- Data imbalance leads to worse performance on uncommon senses



Kilgarriff (2004), *How dominant is the commonest sense of a word?.*
Kumar et al. (2019), *Zero-shot Word Sense Disambiguation using Sense Definition Embeddings.*

Data Sparsity in WSD

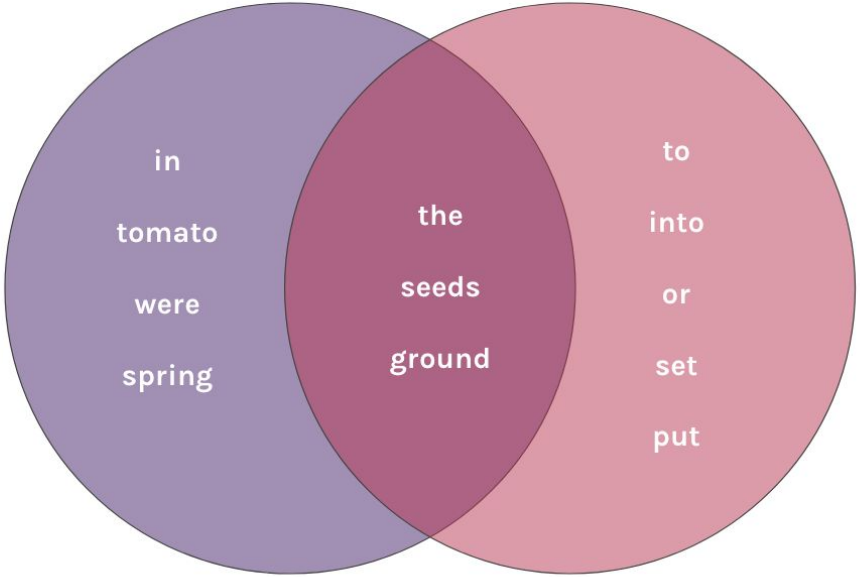
- Senses have Zipfian distribution in natural language
- Data imbalance leads to worse performance on uncommon senses
- We propose an approach to improve performance on rare senses with pretrained models and glosses



Kilgarriff (2004), *How dominant is the commonest sense of a word?.*
Kumar et al. (2019), *Zero-shot Word Sense Disambiguation using Sense Definition Embeddings.*

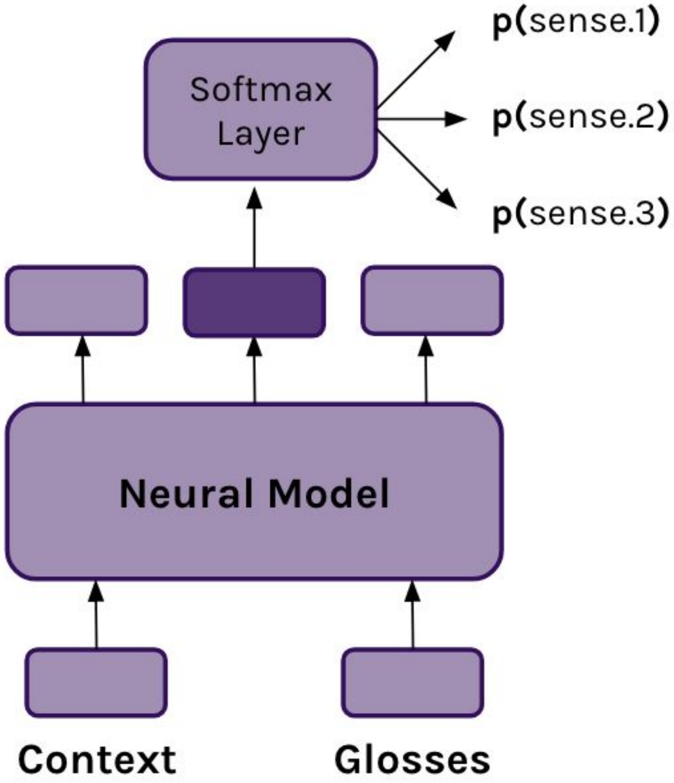
Incorporating Glosses into WSD Models

- Lexical overlap between **context** and **gloss** is a successful knowledge-based approach (Lesk, 1986)



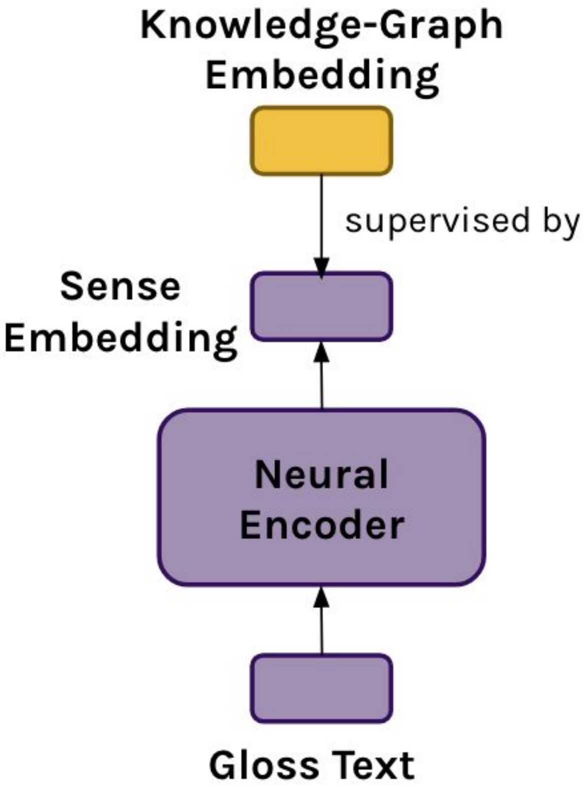
Incorporating Glosses into WSD Models

- Lexical overlap between **context** and **gloss** is a successful knowledge -based approach (Lesk, 1986)
- Neural models integrate glosses by:
  - Adding glosses as additional inputs into the WSD model (Luo et al., 2018a,b)



Incorporating Glosses into WSD Models

- Lexical overlap between **context** and **gloss** is a successful knowledge -based approach (Lesk, 1986)
- Neural models integrate glosses by:
  - Adding glosses as additional inputs into the WSD model (Luo et al., 2018a,b)
  - Mapping encoded gloss representations onto graph embeddings to be used as labels for a WSD model (Kumar et al., 2019)



Pretrained Models for WSD

- Simple **probing** classifiers on frozen pretrained representations found to perform better than models without pretraining

Hadiwinoto et al. (2019), *Improved word sense disambiguation using pretrained contextualized representations.*
Huang et al. (2019), *GlossBERT: Bert for word sense disambiguation with gloss knowledge.*

Pretrained Models for WSD

- Simple **probing** classifiers on frozen pretrained representations found to perform better than models without pretraining
- **GlossBERT** finetunes BERT on WSD with glosses by setting it up as a sentence-pair classification task

Hadiwinoto et al. (2019), *Improved word sense disambiguation using pretrained contextualized representations.*
Huang et al. (2019), *GlossBERT: Bert for word sense disambiguation with gloss knowledge.*

Our Approach: Gloss Informed Bi-encoder

- Two encoders that independently encode the **context** and **gloss**, aligning the target word embedding to the correct sense embedding

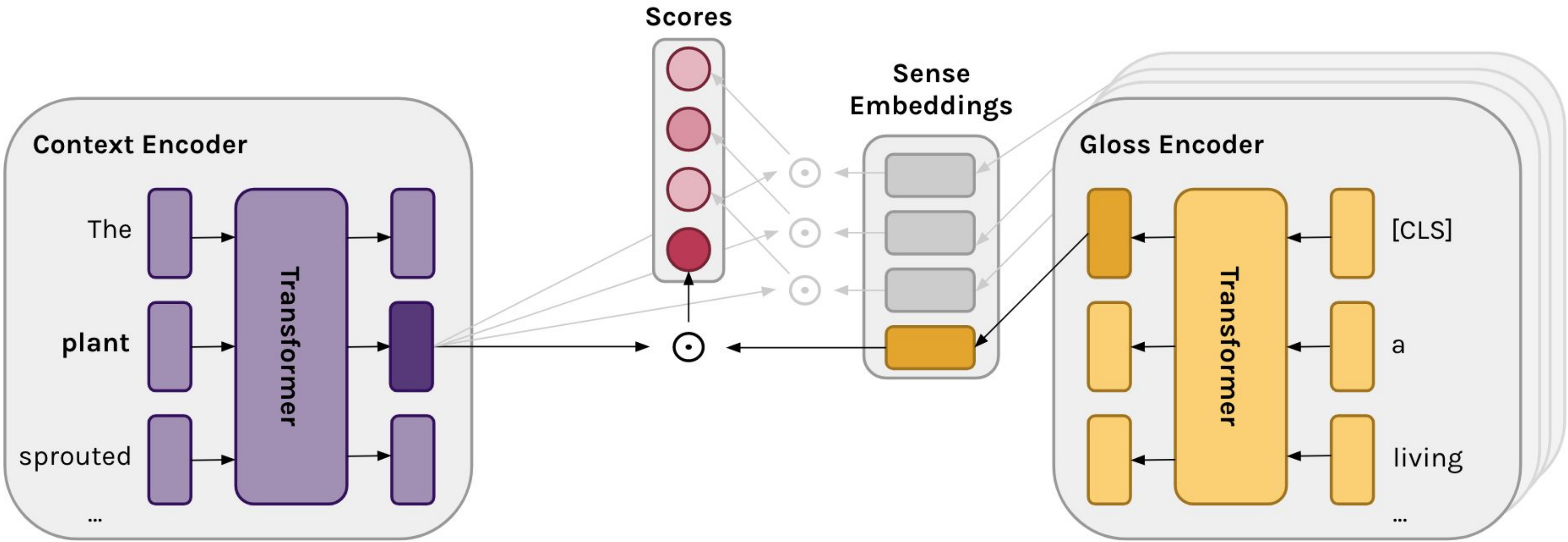
Our Approach: Gloss Informed Bi-encoder

- Two encoders that independently encode the **context** and **gloss**, aligning the target word embedding to the correct sense embedding
- Encoders initialized with BERT and trained end-to-end, without external knowledge

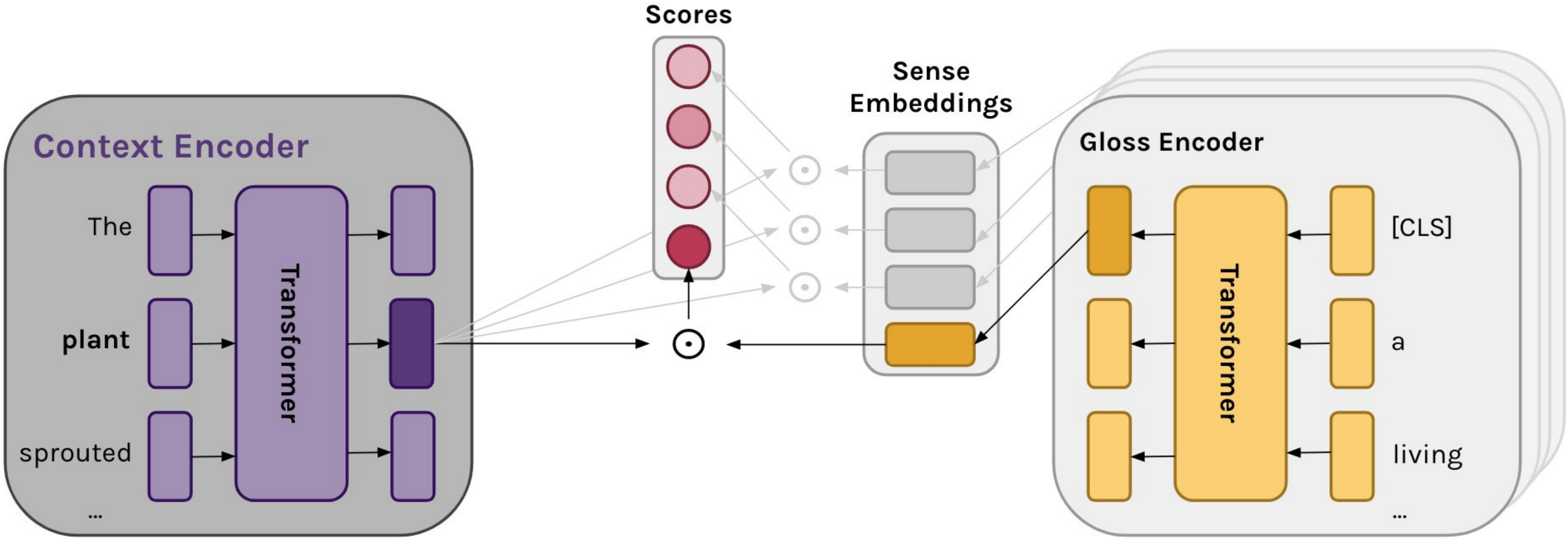
Our Approach: Gloss Informed Bi-encoder

- Two encoders that independently encode the **context** and **gloss**, aligning the target word embedding to the correct sense embedding
- Encoders initialized with BERT and trained end-to-end, without external knowledge
- The bi-encoder is more computationally efficient than a cross-encoder

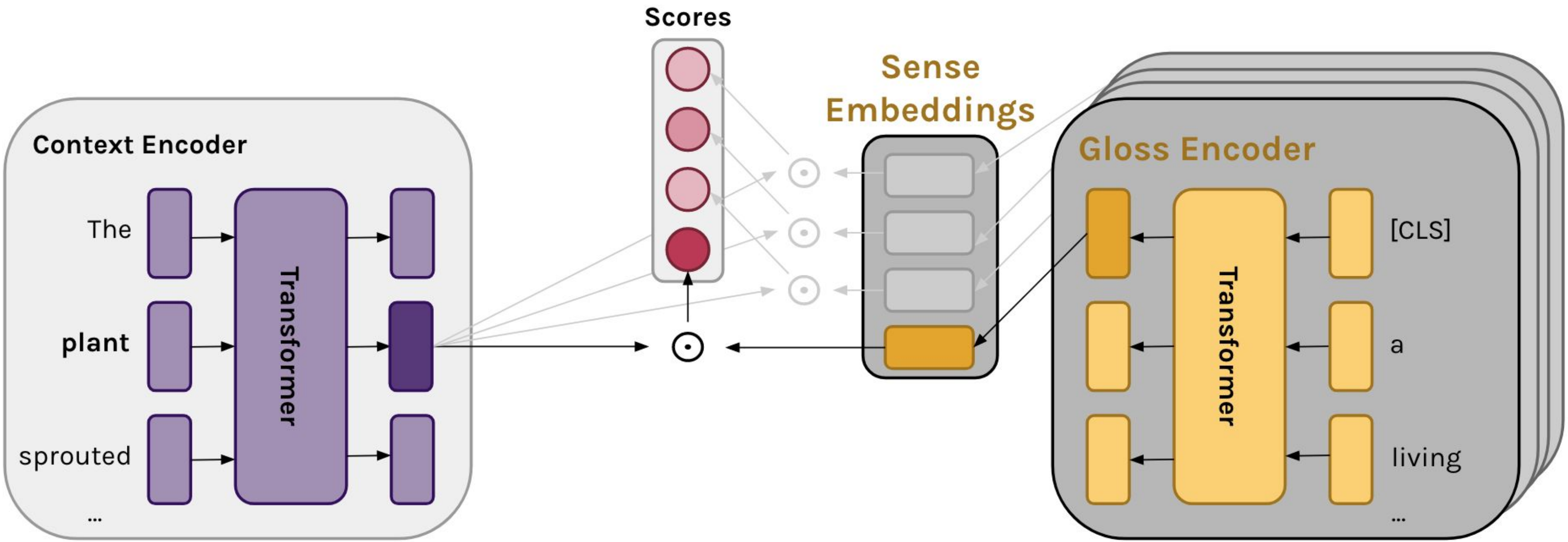
Our Approach: Gloss Informed Bi-encoder



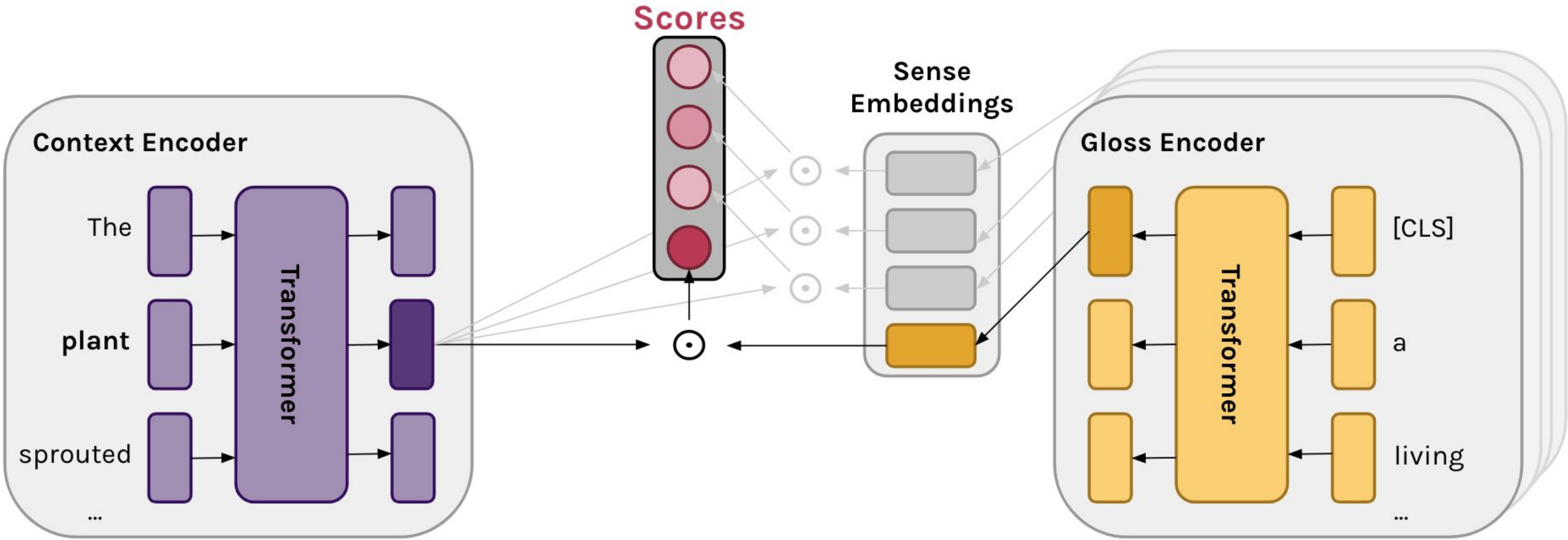
Our Approach: Gloss Informed Bi-encoder



Our Approach: Gloss Informed Bi-encoder



Our Approach: Gloss Informed Bi-encoder



Baselines and Prior Work

Model	Glosses?	Pretraining?	Source
HCAN	✓		Luo et al., 2018a
EWISE	✓		Kumar et al., 2019
BERT **Probe**		✓	**Ours**
GLU		✓	Hadiwinoto et al., 2019
LMMS	✓	✓	Loureiro and Jorge, 2019
SVC		✓	Vial et al., 2019
GlossBERT	✓	✓	Huang et al., 2019
Bi-encoder Model (**BEM**)	✓	✓	**Ours**

Baselines and Prior Work

Model	Glosses?	Pretraining?	Source
HCAN	✓		Luo et al., 2018a
EWISE	✓		Kumar et al., 2019
BERT **Probe**		✓	**Ours**
GLU		✓	Hadiwinoto et al., 2019
LMMS	✓	✓	Loureiro and Jorge, 2019
SVC		✓	Vial et al., 2019
GlossBERT	✓	✓	Huang et al., 2019
Bi-encoder Model (**BEM**)	✓	✓	**Ours**

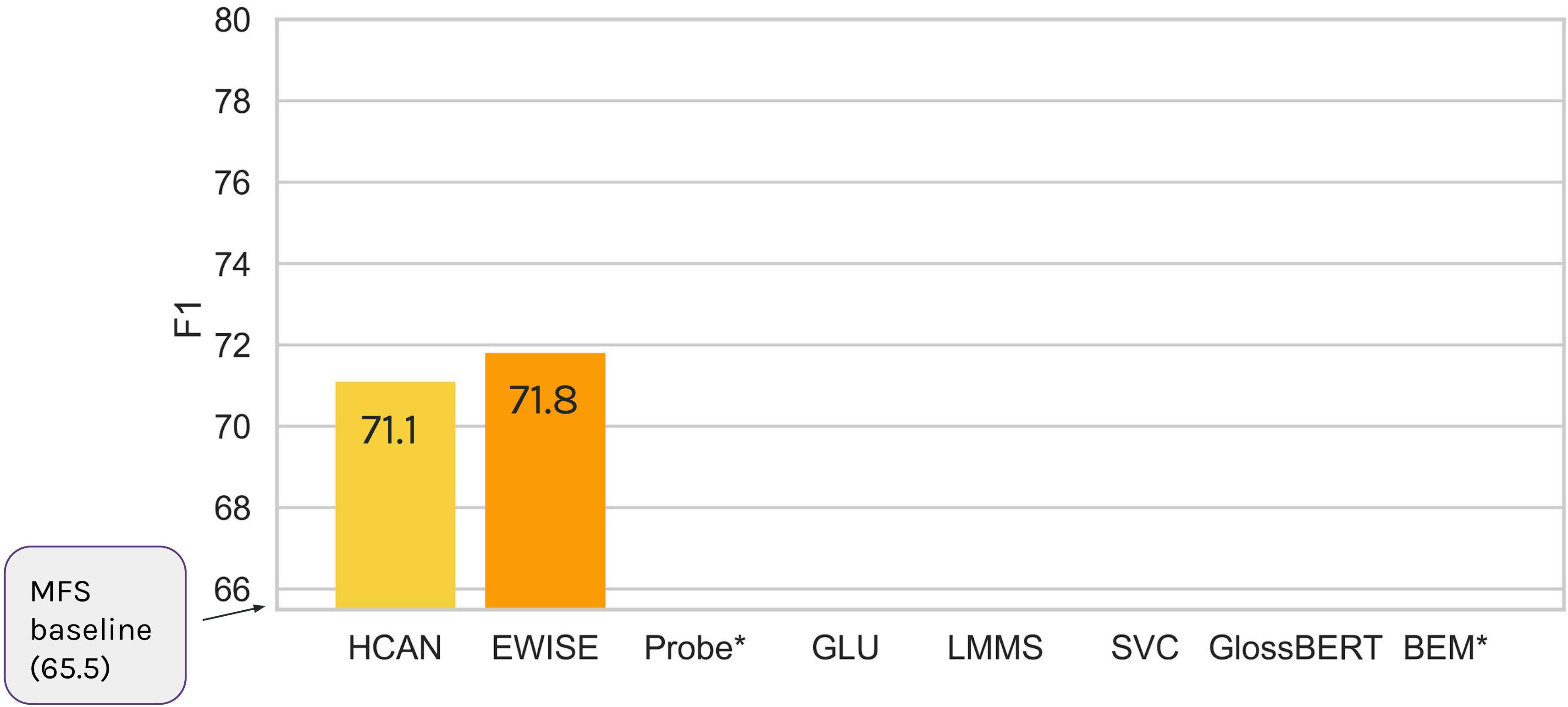
Baselines and Prior Work

Model	Glosses?	Pretraining?	Source
HCAN	✓		Luo et al., 2018a
EWISE	✓		Kumar et al., 2019
BERT **Probe**		✓	**Ours**
GLU		✓	Hadiwinoto et al., 2019
LMMS	✓	✓	Loureiro and Jorge, 2019
SVC		✓	Vial et al., 2019
GlossBERT	✓	✓	Huang et al., 2019
Bi-encoder Model (**BEM**)	✓	✓	**Ours**

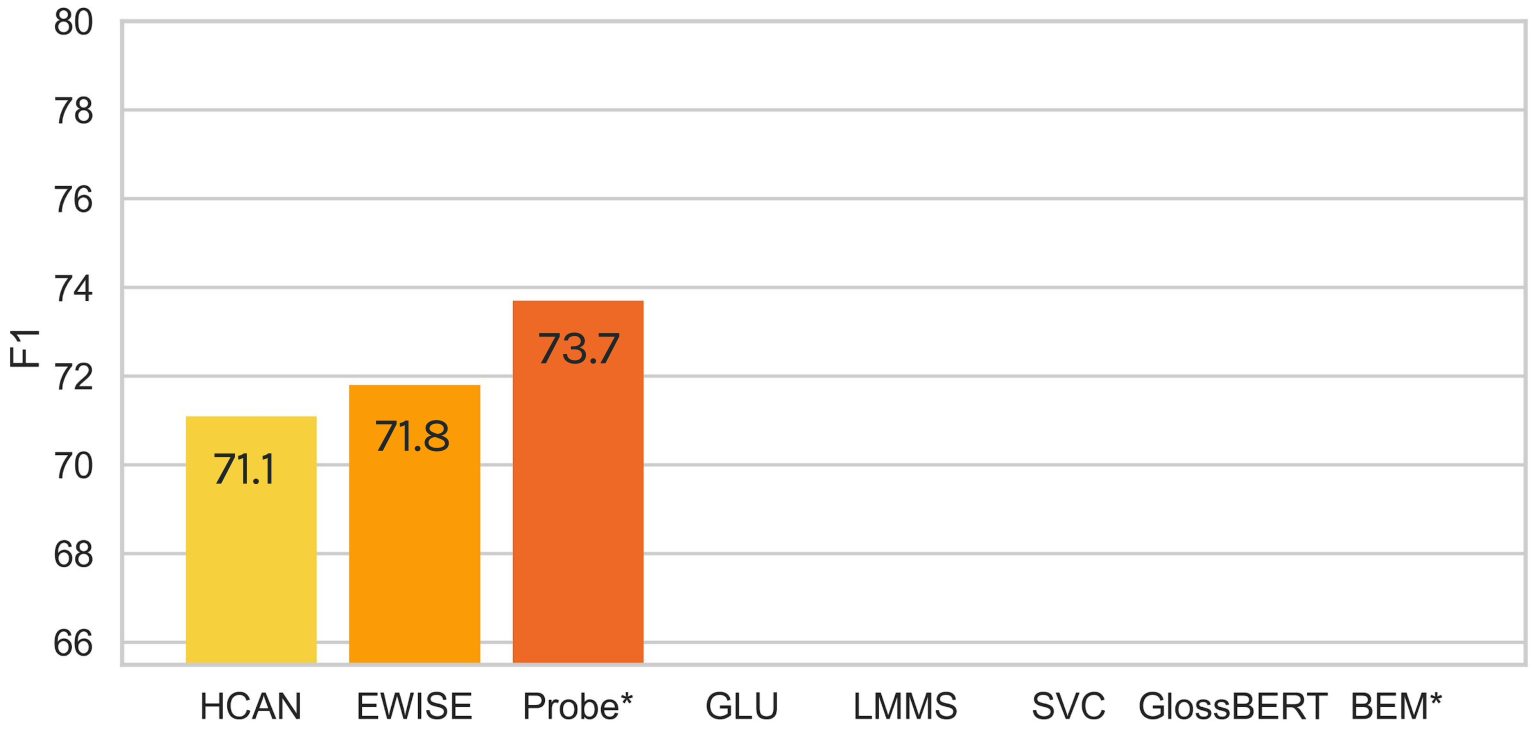
Baselines and Prior Work

Model	Glosses?	Pretraining?	Source
HCAN	✓		Luo et al., 2018a
EWISE	✓		Kumar et al., 2019
BERT **Probe**		✓	**Ours**
GLU		✓	Hadiwinoto et al., 2019
LMMS	✓	✓	Loureiro and Jorge, 2019
SVC		✓	Vial et al., 2019
GlossBERT	✓	✓	Huang et al., 2019
Bi-encoder Model (**BEM**)	✓	✓	**Ours**

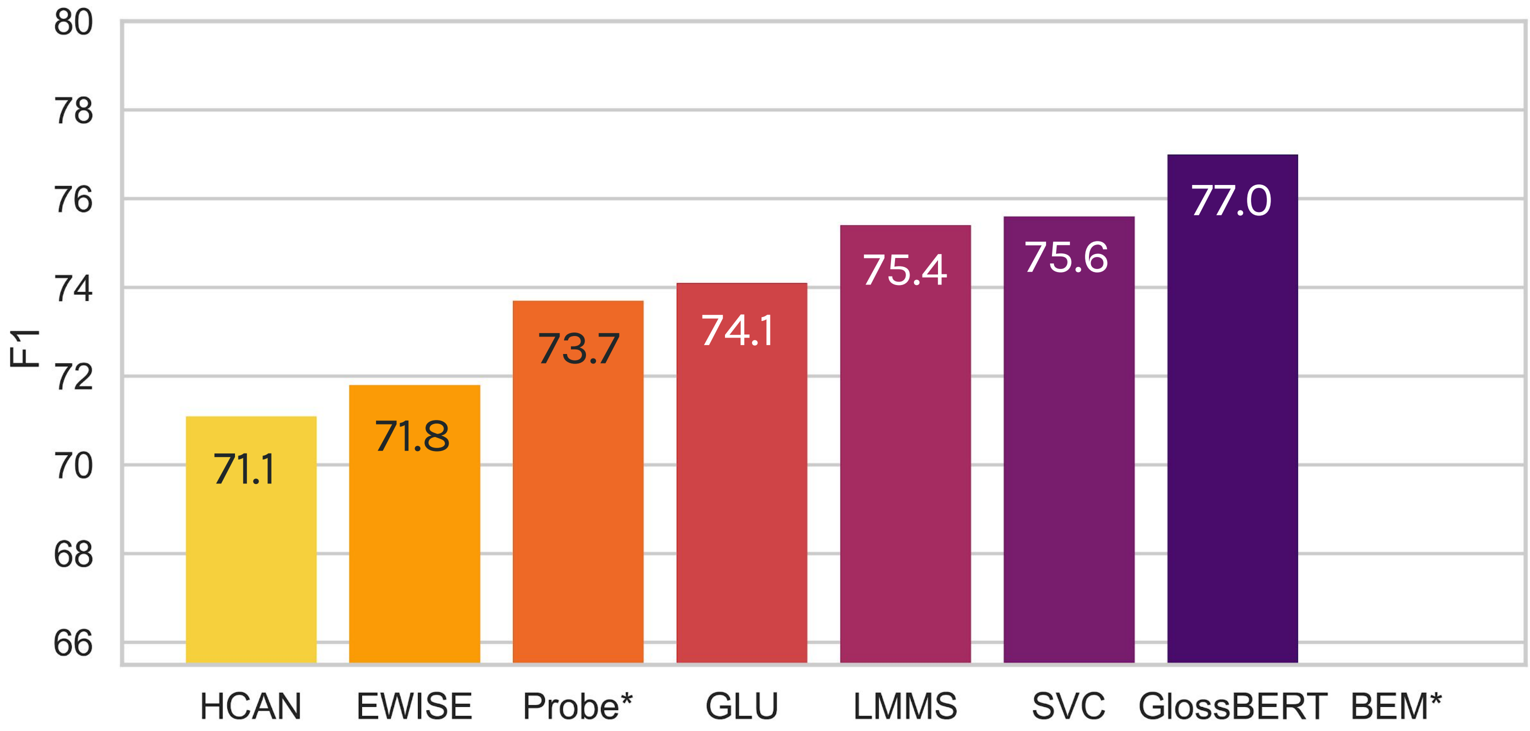
Overall WSD Performance



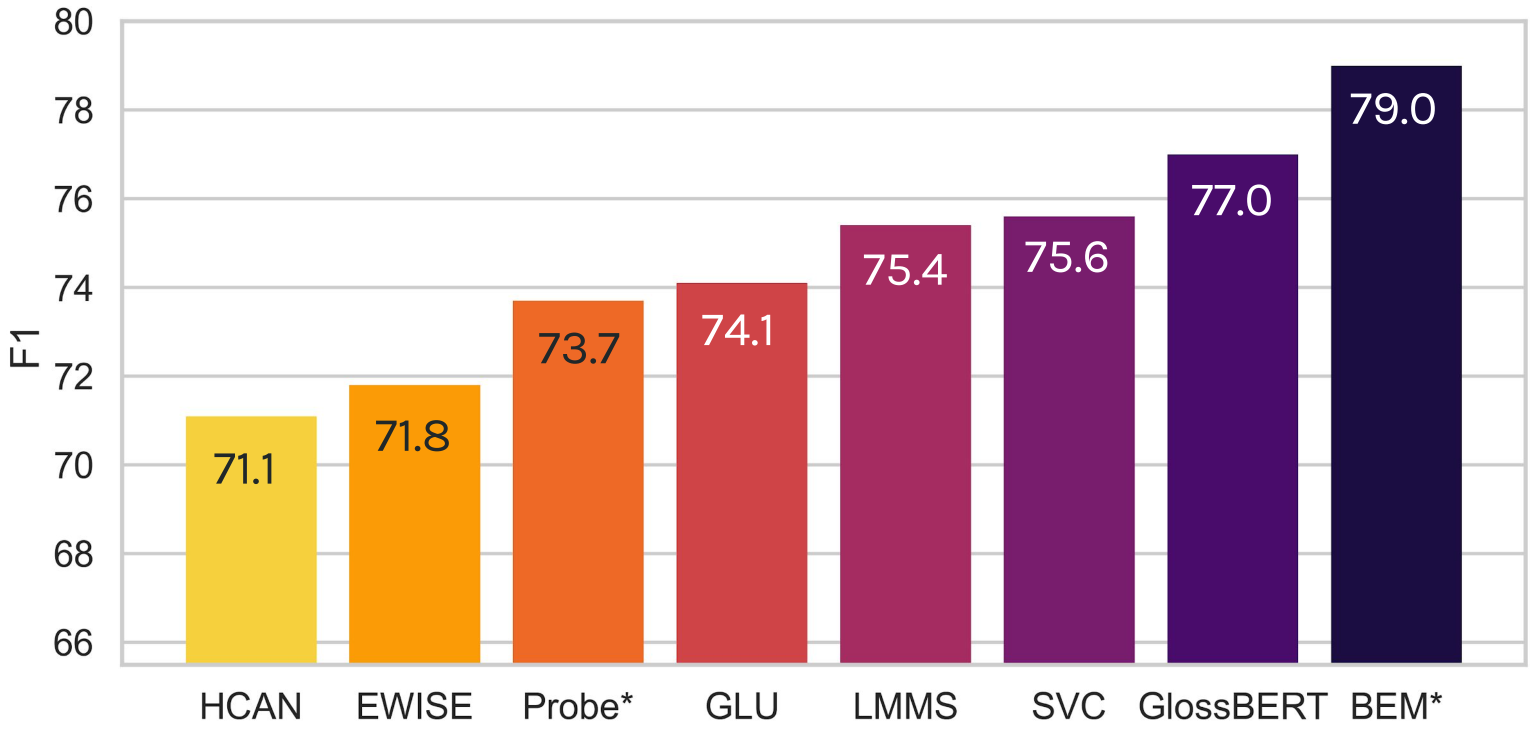
Overall WSD Performance



Overall WSD Performance

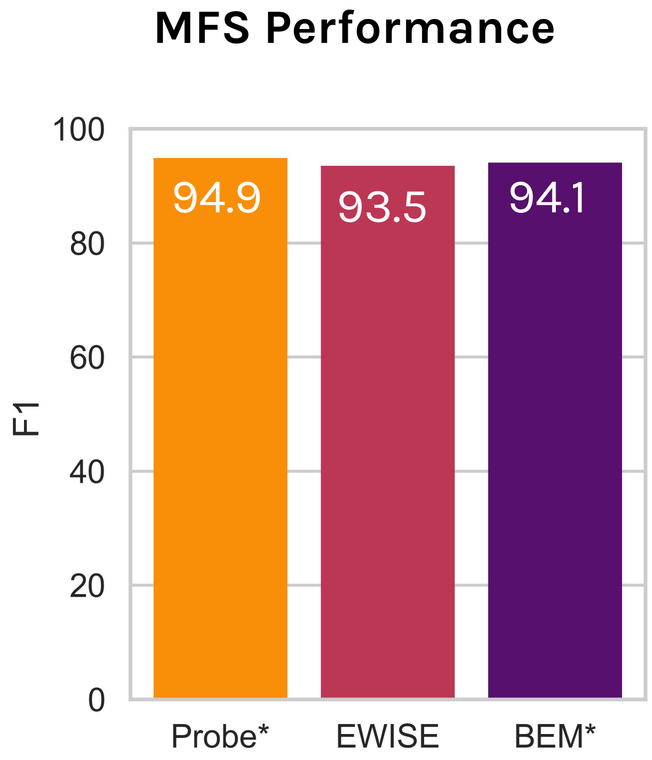


Overall WSD Performance

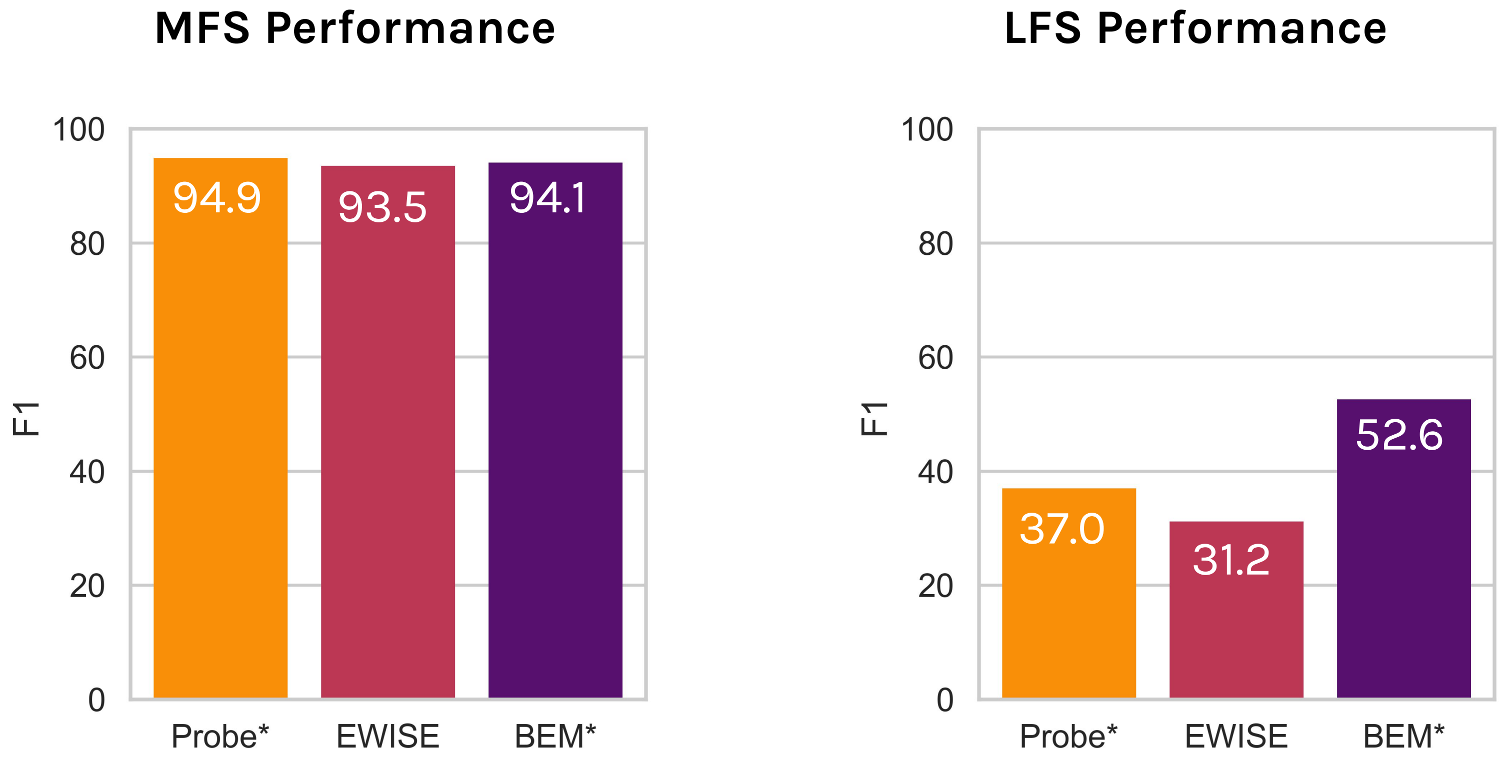


Performance by Sense Frequency

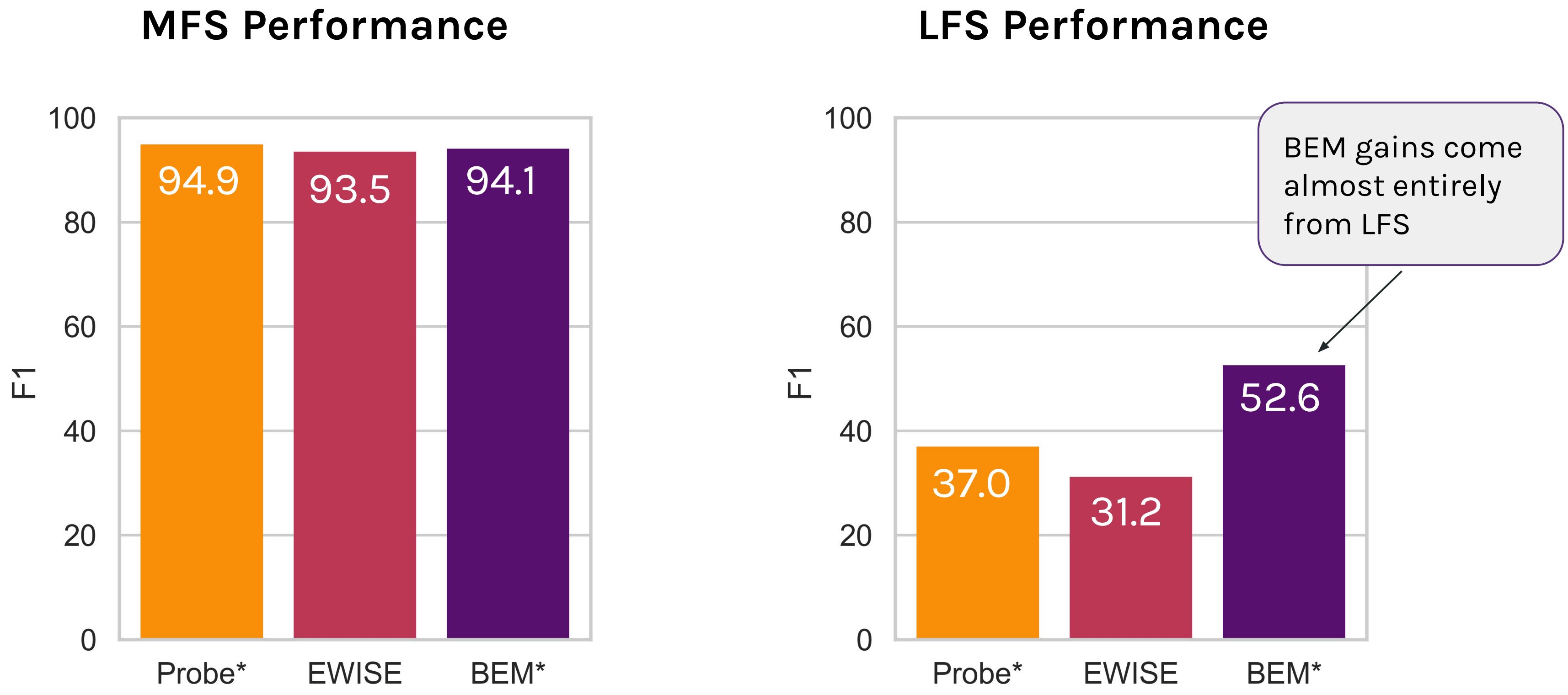
Performance by Sense Frequency



Performance by Sense Frequency



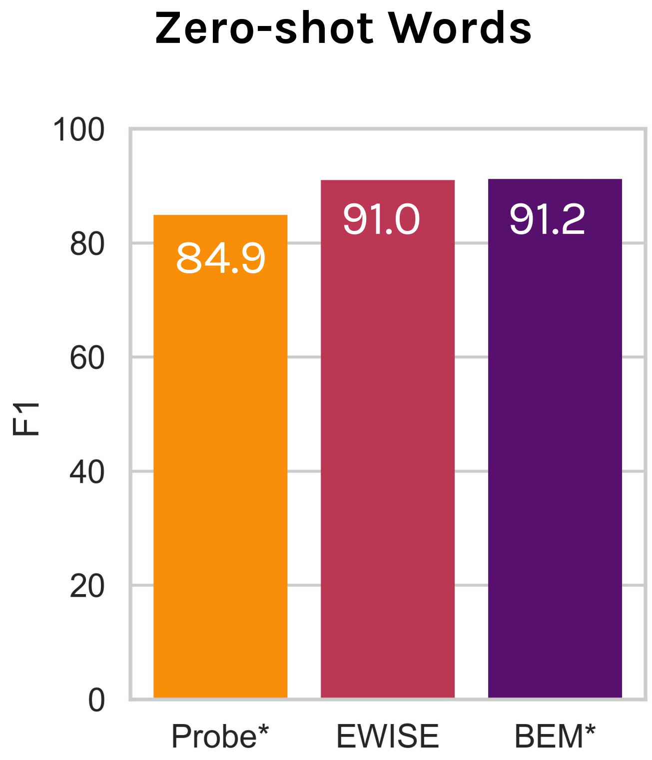
Performance by Sense Frequency



Zero-shot Evaluation

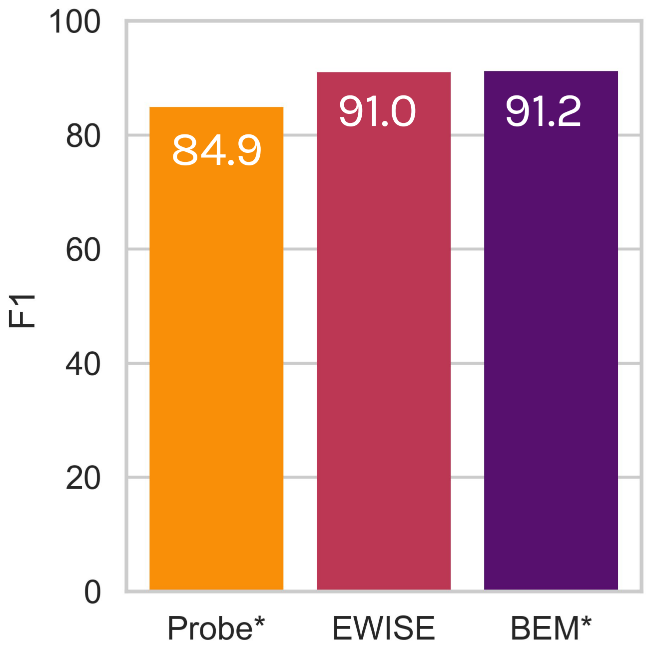
- BEM can represent new, unseen senses with gloss encoder and encode unseen words with the context encoder
- Probe baseline relies on **WordNet back-off**, predicting the most common sense of unseen words as indicated in WordNet

Zero-shot Evaluation

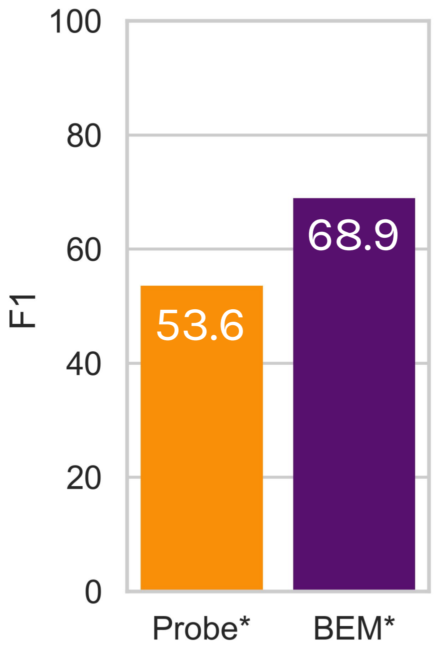


Zero-shot Evaluation

Zero-shot Words



Zero-shot Senses

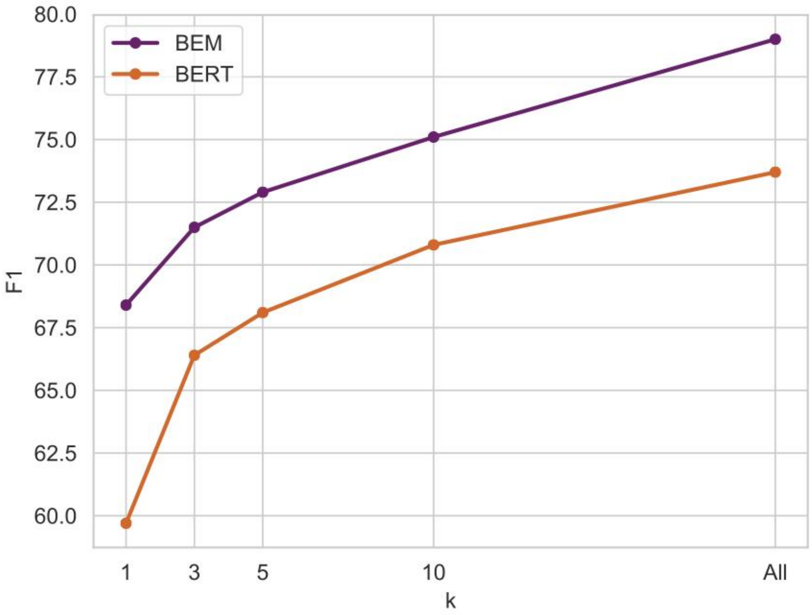


Few-shot Learning of WSD

Train BEM (and frozen probe baseline) on subset of SemCor, with (up to) **k** examples of each sense in the training data

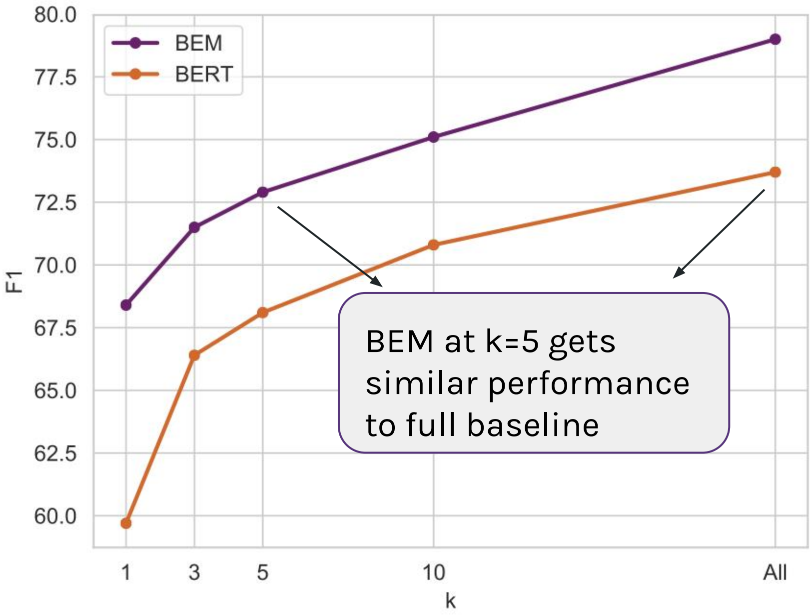
Few-shot Learning of WSD

Train BEM (and frozen probe baseline) on subset of SemCor, with (up to) **k** examples of each sense in the training data



Few-shot Learning of WSD

Train BEM (and frozen probe baseline) on subset of SemCor, with (up to) **k** examples of each sense in the training data



Takeaways

- The **BEM** improves over the BERT probe baseline and prior approaches to using (1) sense definitions and (2) pretrained models for WSD

Takeaways

- The **BEM** improves over the BERT probe baseline and prior approaches for using (1) sense definitions and (2) pretrained models for WSD
- Gains stem from better performance on less common and unseen senses

Takeaways

- The **BEM** improves over the BERT probe baseline and prior approaches to using (1) sense definitions and (2) pretrained models for WSD
- Gains stem from better performance on less common and unseen senses

Questions?

https://github.com/facebookresearch/wsd-biencoders

blvns@cs.washington.edu