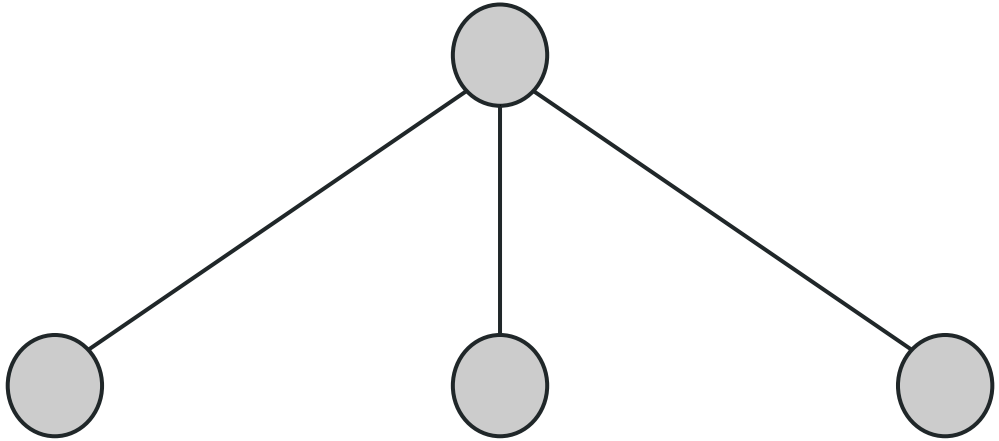
FEWS: Large-Scale, Low-Shot Word Sense Disambiguation with the Dictionary

Terra Blevins, Mandar Joshi, and Luke Zettlemoyer

facebook
AI Research

I **liked** my friend’s status.

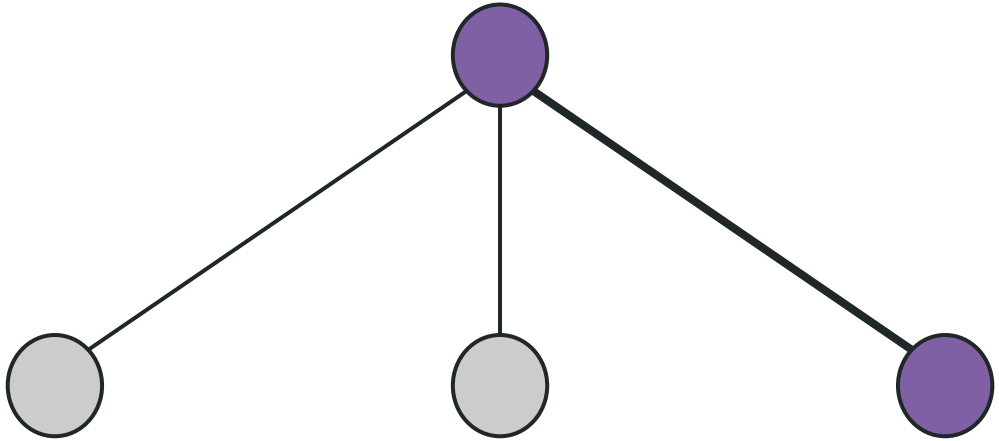


(v) To enjoy... [or] be in favor of.

(v) To find attractive; to prefer the company of.

(v) To show support for something on the Internet...

I **liked** my friend’s status.



(v) To enjoy... [or] be in favor of.

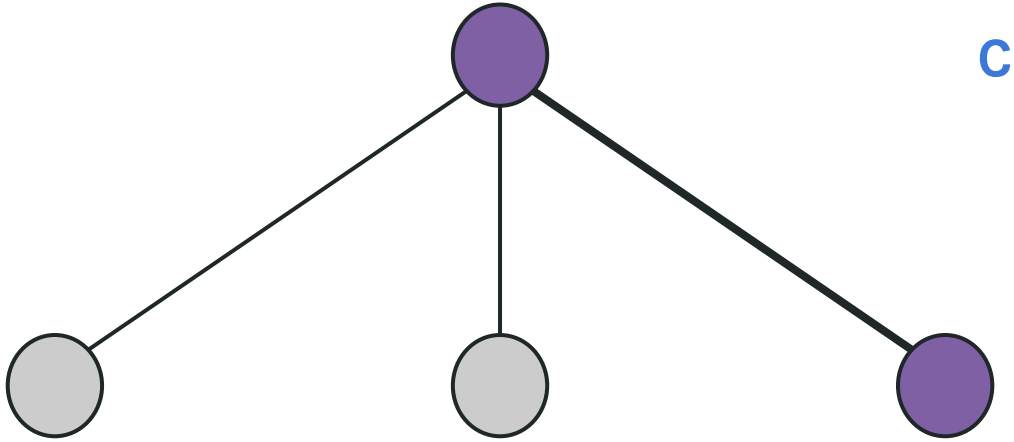
(v) To find attractive; to prefer the company of.

(v) To show support for something on the Internet...

Target Word

I **liked** my friend's status.

Context



(v) To enjoy... [or] be in favor of.

(v) To find attractive; to prefer the company of.

(v) To show support for something on the Internet...

Candidate Senses

Data Sparsity in WSD

- Senses have Zipfian distribution in natural language text

Kilgarriff (2004), *How dominant is the commonest sense of a word?*.
Miller et al. (1993). *A Semantic correspondence.*

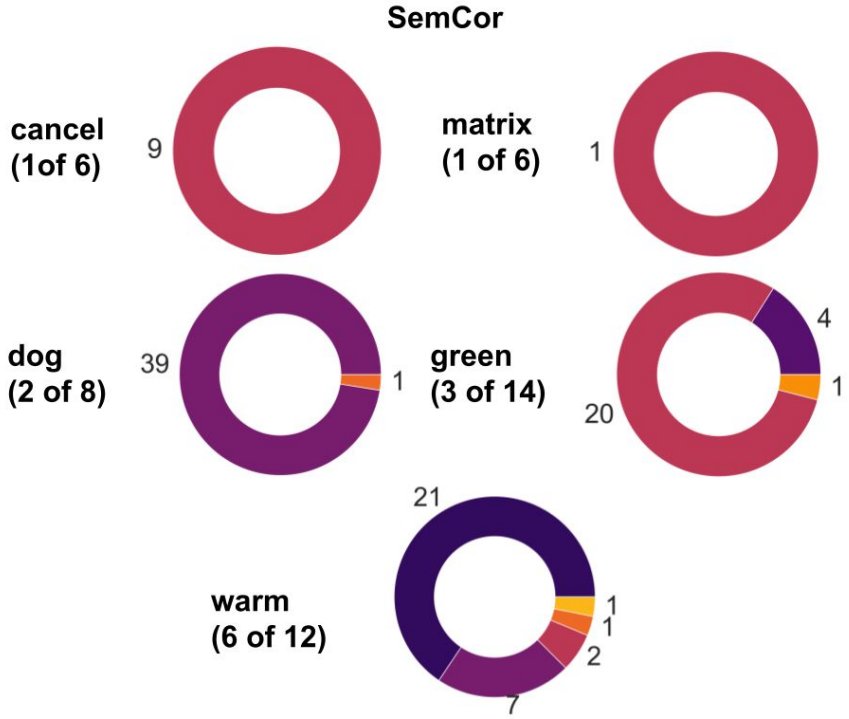
Data Sparsity in WSD

- Senses have Zipfian distribution in natural language text
- Data imbalance leads to fewer examples for uncommon senses

Kilgarriff (2004), *How dominant is the commonest sense of a word?*.
Miller et al. (1993). *A Semantic correspondence.*

Data Sparsity in WSD

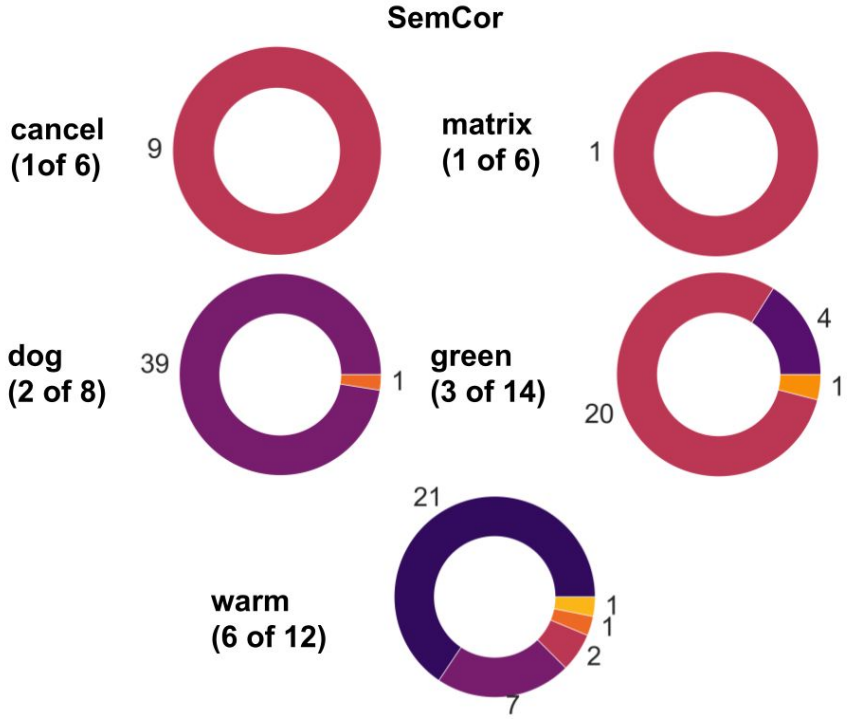
- Senses have Zipfian distribution in natural language text
- Data imbalance leads to fewer examples for uncommon senses



Kilgarriff (2004), *How dominant is the commonest sense of a word?*.
Miller et al. (1993). *A Semantic correspondence.*

Data Sparsity in WSD

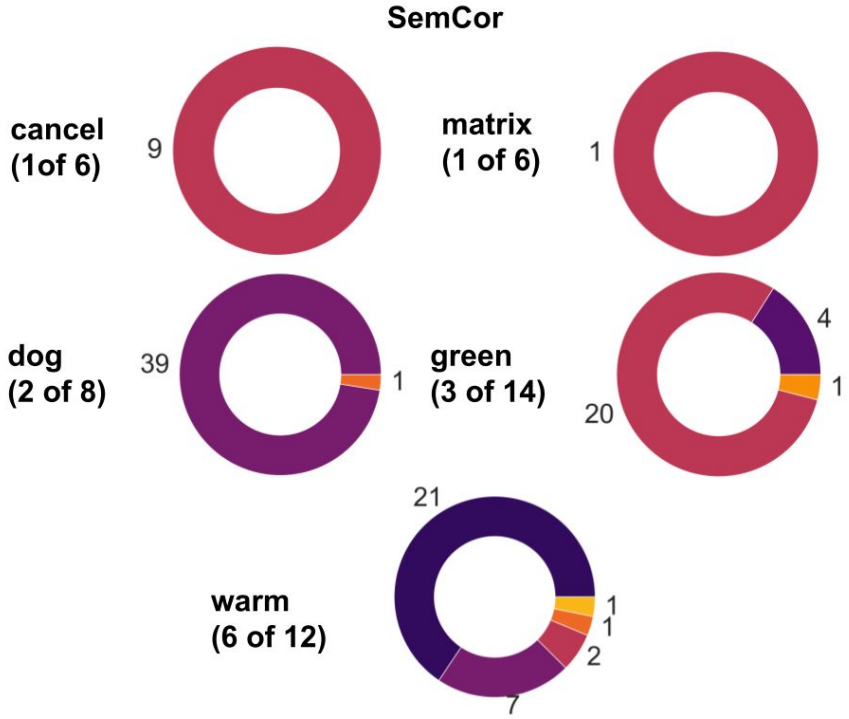
- Senses have Zipfian distribution in natural language text
- Data imbalance leads to fewer examples for uncommon senses
- This leads to:
  - (Very) limited training data for rare senses



Kilgarriff (2004), *How dominant is the commonest sense of a word?*.
Miller et al. (1993). *A Semantic correspondence.*

Data Sparsity in WSD

- Senses have Zipfian distribution in natural language text
- Data imbalance leads to fewer examples for uncommon senses
- This leads to:
  - (Very) limited training data for rare senses
  - Unreliable evaluation of model performance on rare senses



Kilgarriff (2004), *How dominant is the commonest sense of a word?.*
Miller et al. (1993). *A Semantic correspondence.*

Few-shot Examples of Word Sense (FEWS)

- To address the data sparsity issue for rare senses, we create **FEWS**, a new WSD dataset

Few-shot Examples of Word Sense (FEWS)

- To address the data sparsity issue for rare senses, we create **FEWS**, a new WSD dataset
- Data in FEWS come from Wiktionary example sentences

Few-shot Examples of Word Sense (FEWS)

- To address the data sparsity issue for rare senses, we create **FEWS**, a new WSD dataset
- Data in FEWS come from Wiktionary example sentences
- Using a **dictionary** as a data source means that FEWS is:
  - High coverage (particularly on rare senses)
  - Low-shot (only a few labeled examples per sense)

Few-shot Examples of Word Sense (FEWS)

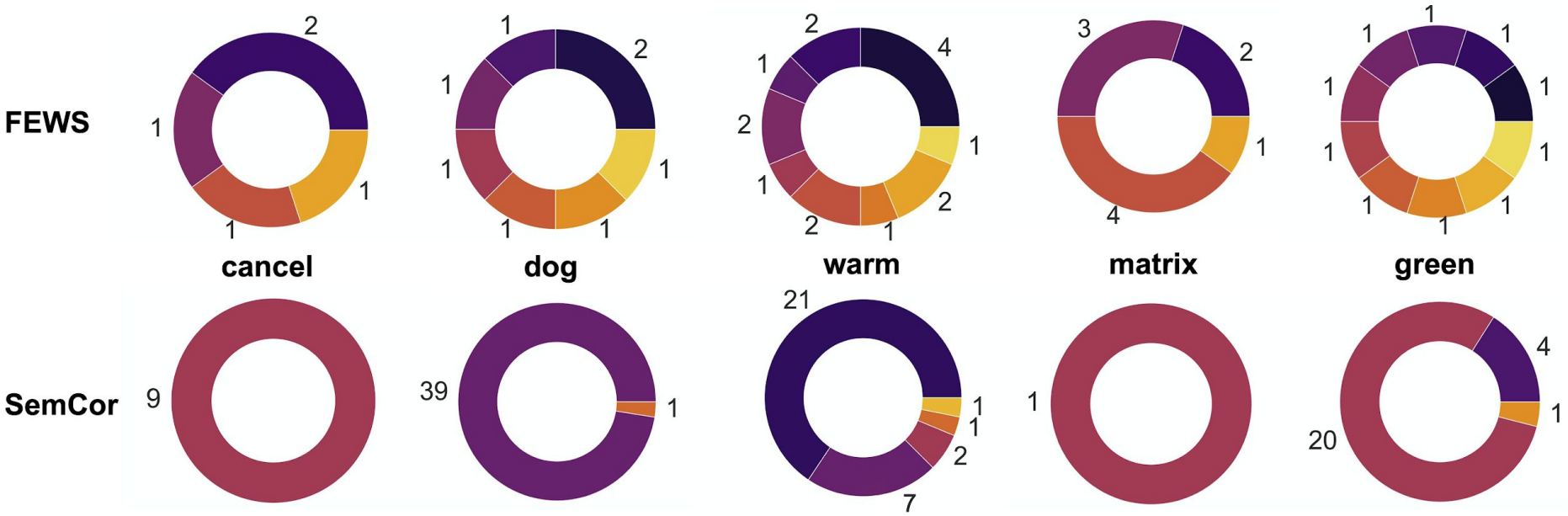
- FEWS consists of a **glossary** of word senses and their definitions, a **training set** (121k examples) and development and test **evaluation sets** (10k examples each).

Few-shot Examples of Word Sense (FEWS)

- FEWS consists of a **glossary** of word senses and their definitions, a **training set** (121k examples) and development and test **evaluation sets** (10k examples each).
- The evaluation sets are each split up into **few-shot** and **zero-shot** evaluation settings

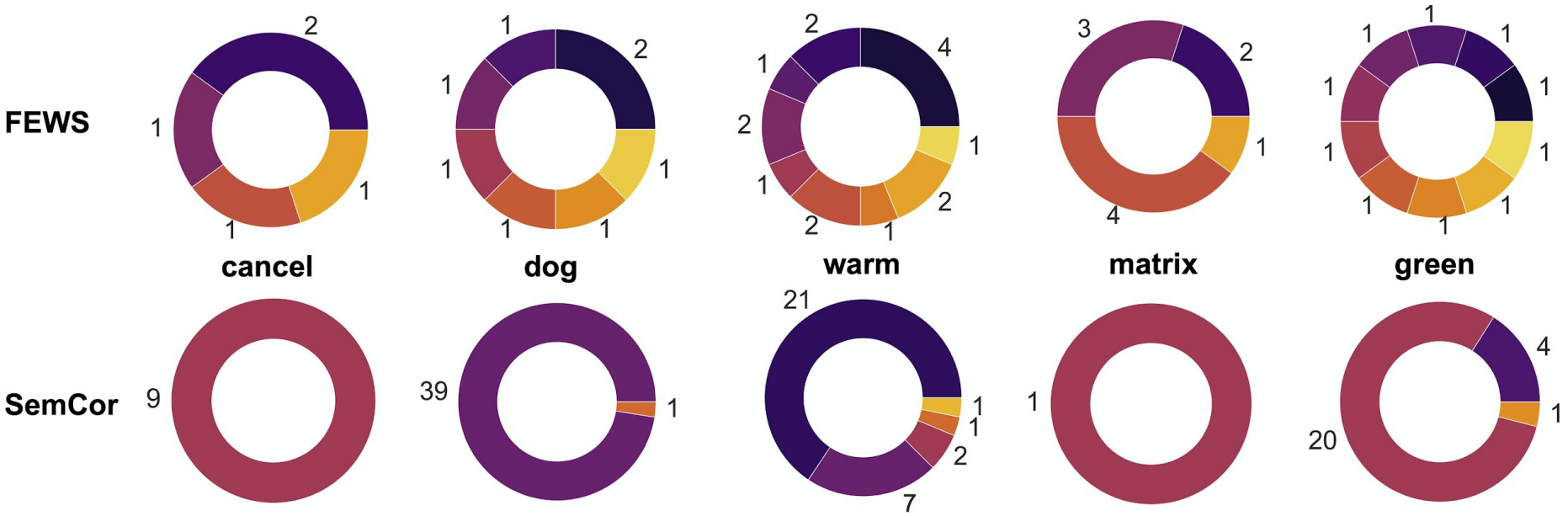
Dataset Analysis of FEWS

- FEWS is a high coverage dataset.



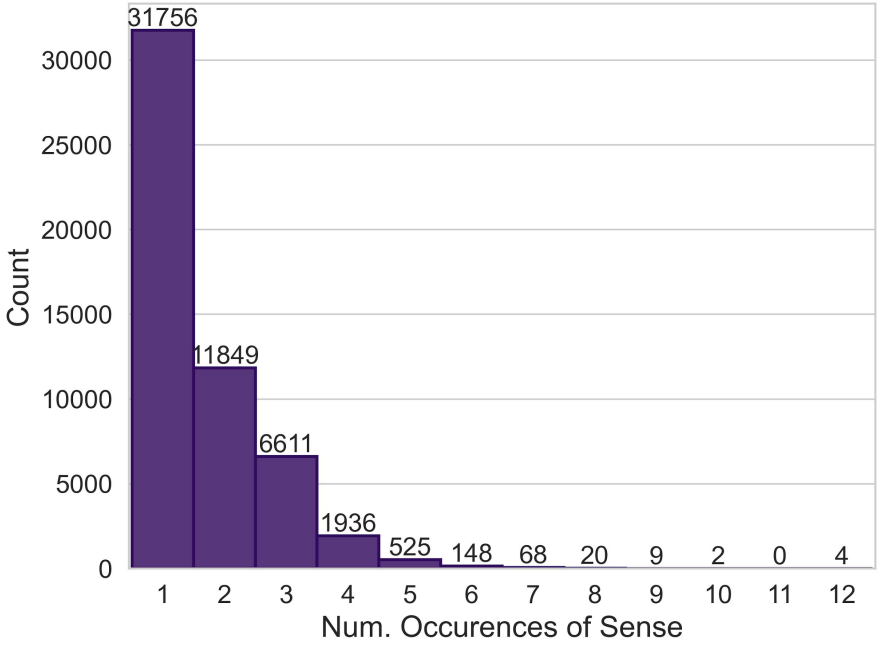
Dataset Analysis of FEWS

- FEWS is a high coverage...
- ... and low-shot dataset.



Dataset Analysis of FEWS

- FEWS is a high coverage...
- ... and low-shot dataset.



Dataset Analysis of FEWS

- FEWS is a high coverage...
- ... and low-shot dataset.
- FEWS also covers a wide range of domains.

Baselines for FEWS

Baselines for FEWS

Baseline	Knowledge-based?	Neural?	Source
Most Frequent Sense (MFS)	✓		Kilgarriff, 2004
Lesk	✓		Kilgarriff and Rosenzweig, 2000
Lesk+Embed	✓		Basile et al., 2014
BERT **Probe**		✓	Blevins and Zettlemoyer, 2020
Bi-encoder Model (**BEM**)	✓	✓	Blevins and Zettlemoyer, 2020
(Est.) **Human Performance**			**Ours**

Baselines for FEWS

Knowledge-based: (usually) untrained baselines that predict word sense based on features of the dataset (i.e., global statistics, glosses)

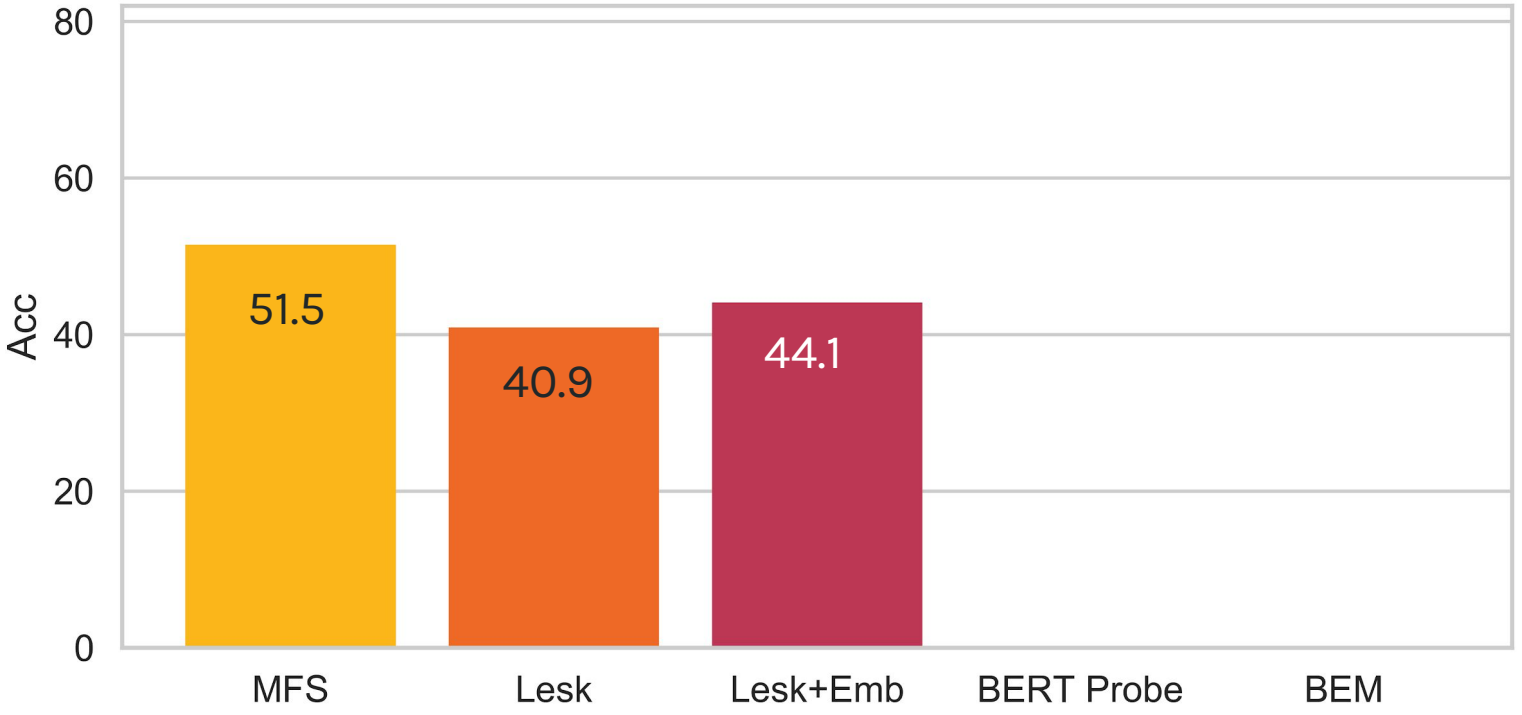
Baseline	Knowledge-based?	Neural?	Source
Most Frequent Sense (MFS)	✓		Kilgarriff, 2004
Lesk	✓		Kilgarriff and Rosenzweig, 2000
Lesk+Embed	✓		Basile et al., 2014
BERT **Probe**		✓	Blevins and Zettlemoyer, 2020
Bi-encoder Model (**BEM**)	✓	✓	Blevins and Zettlemoyer, 2020
(Est.) **Human Performance**			**Ours**

Baselines for FEWS

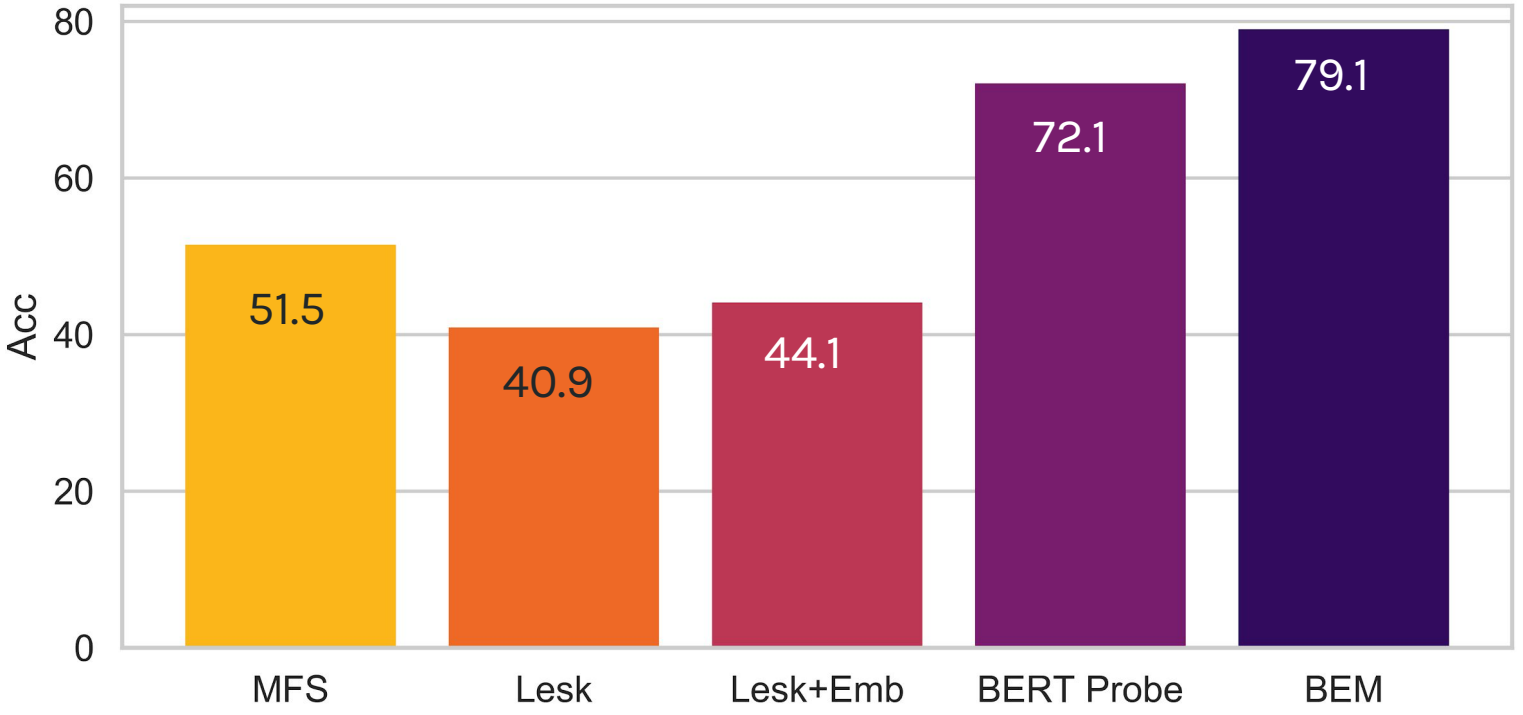
Neural: machine learning baselines that build on pretrained encoders with transformer architectures (BERT)

Baseline	Knowledge-based?	Neural?	Source
Most Frequent Sense (MFS)	✓		Kilgarriff, 2004
Lesk	✓		Kilgarriff and Rosenzweig, 2000
Lesk+Embed	✓		Basile et al., 2014
BERT **Probe**		✓	Blevins and Zettlemoyer, 2020
Bi-encoder Model (**BEM**)	✓	✓	Blevins and Zettlemoyer, 2020
(Est.) **Human Performance**			**Ours**

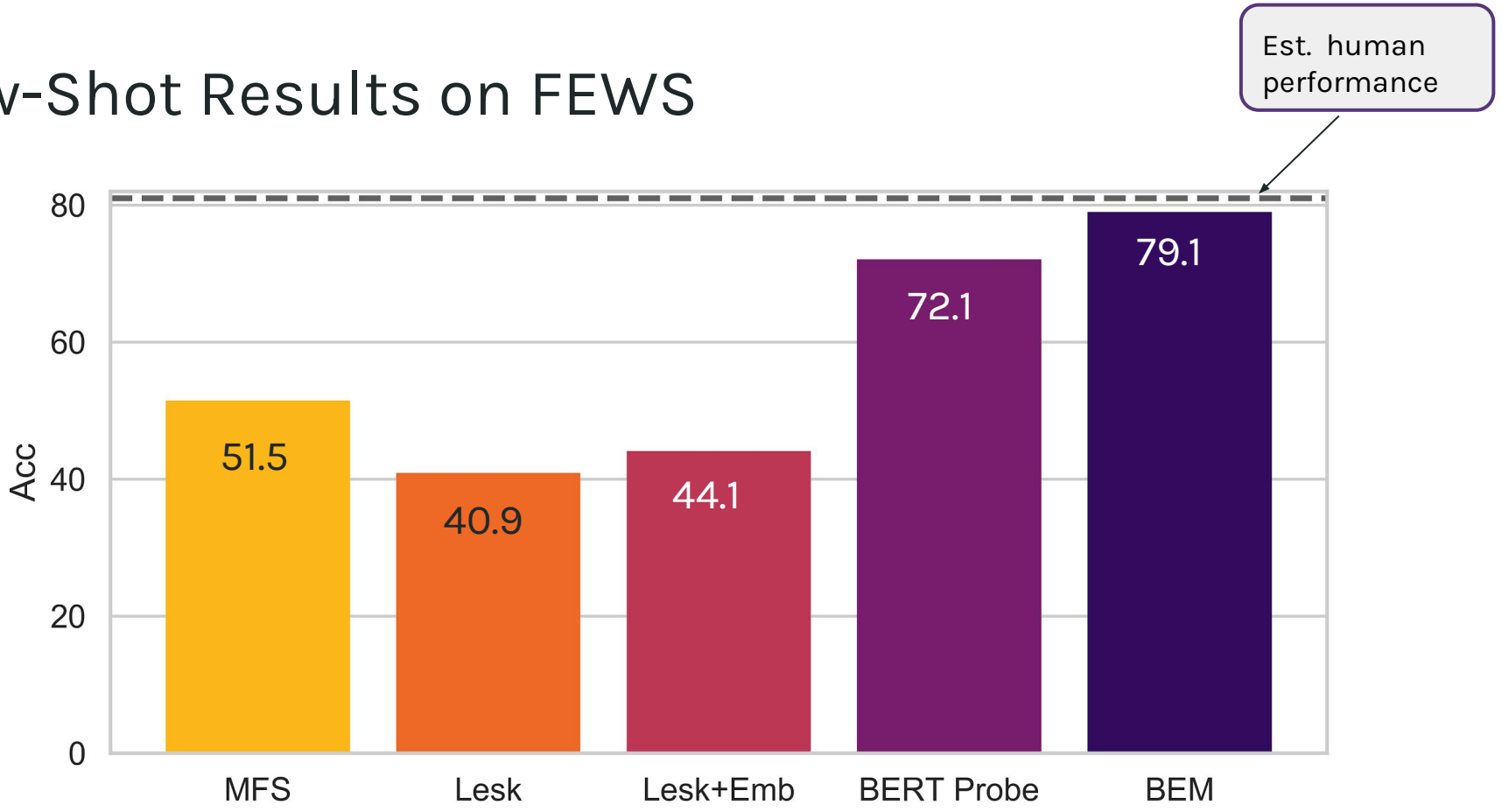
Few-Shot Results on FEWS



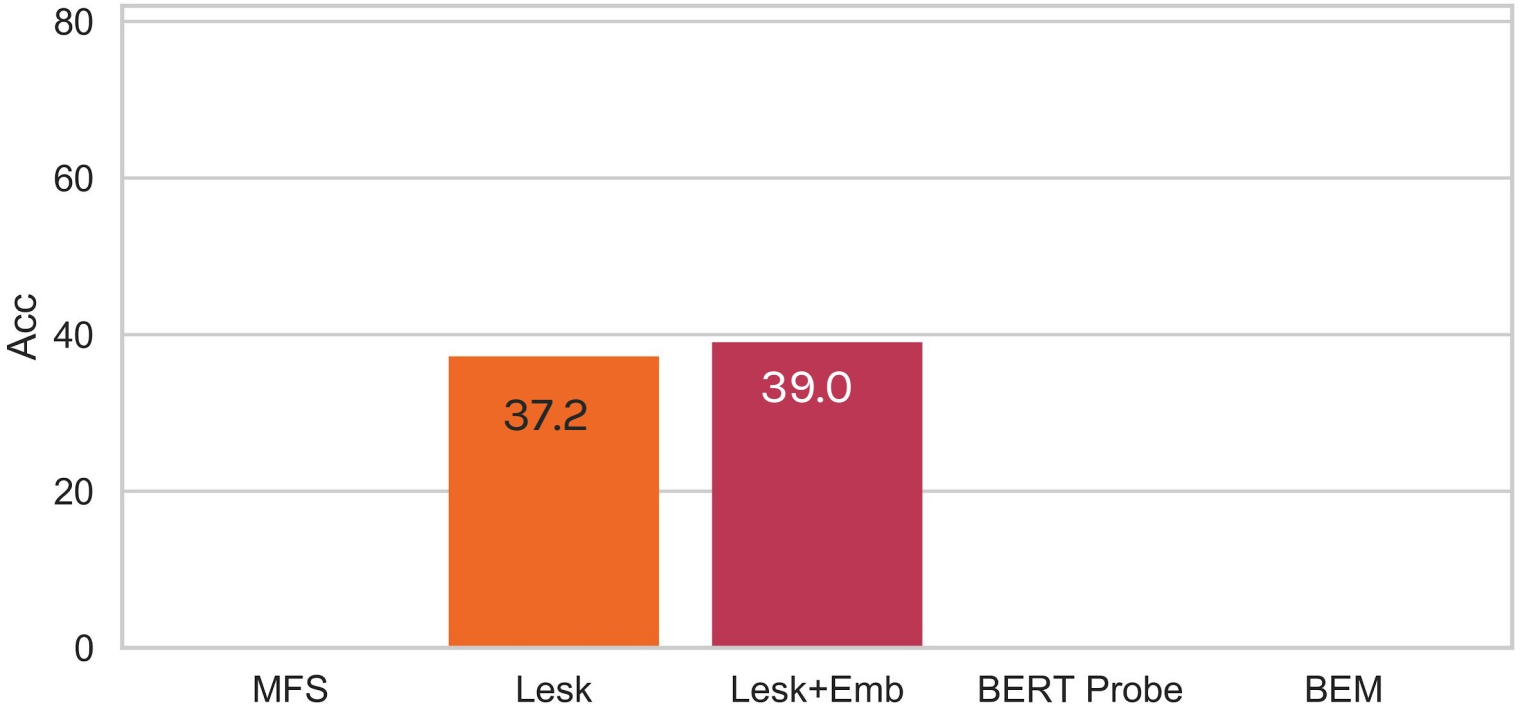
Few-Shot Results on FEWS



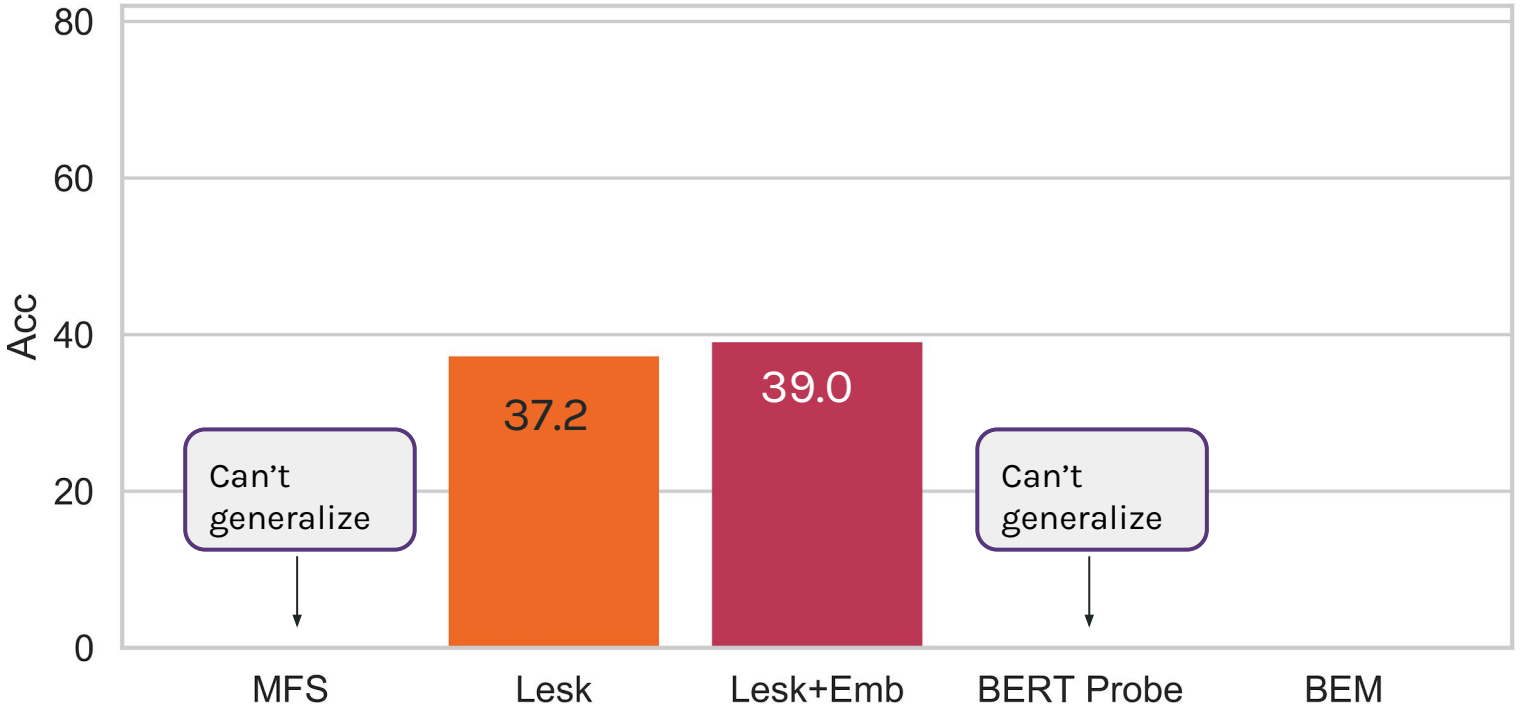
Few-Shot Results on FEWS



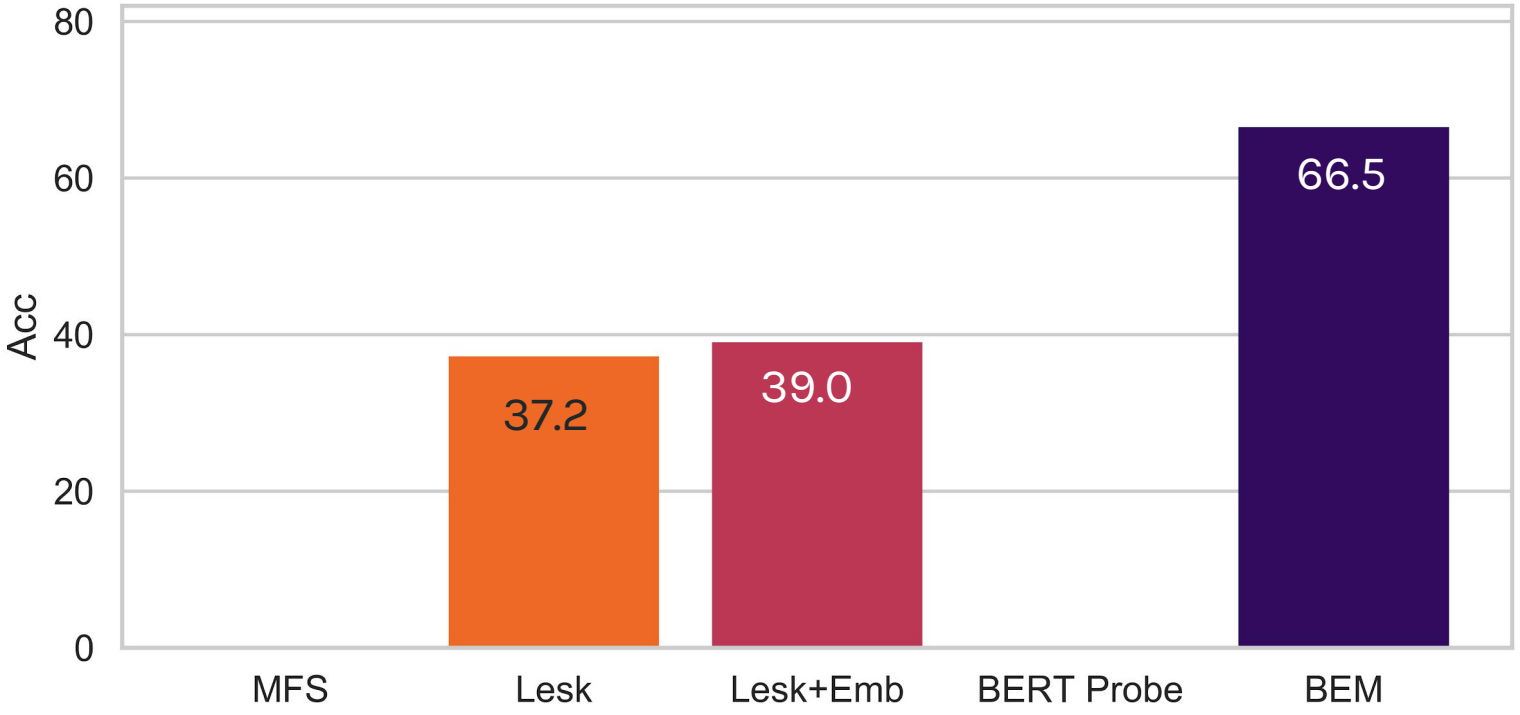
Zero-Shot Results on FEWS



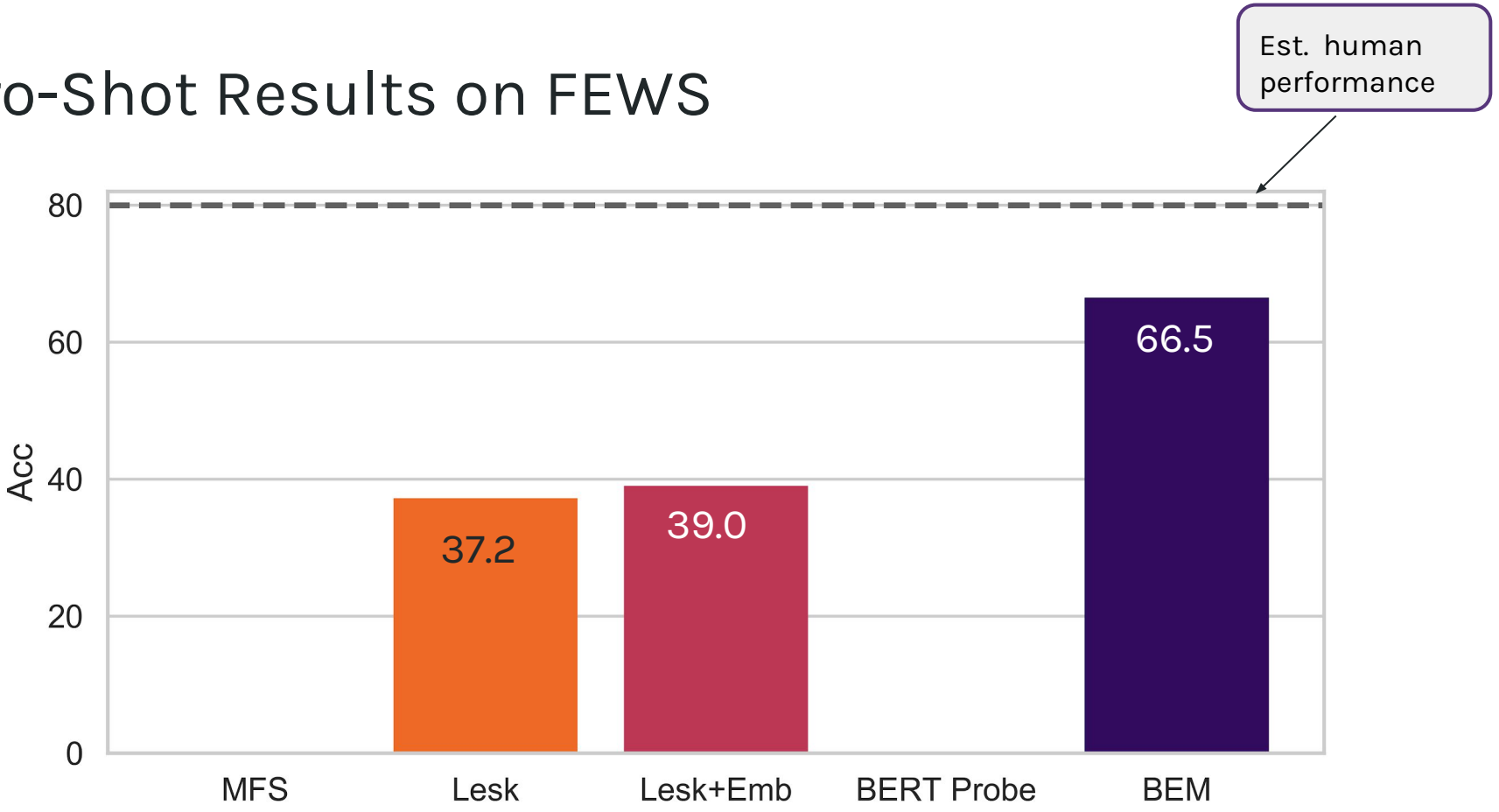
Zero-Shot Results on FEWS



Zero-Shot Results on FEWS



Zero-Shot Results on FEWS



Transfer Learning With FEWS

Transfer Learning With FEWS

- Experiments to evaluate whether FEWS improves performance on uncommon senses in other WSD datasets

Transfer Learning With FEWS

- Experiments to evaluate whether FEWS improves performance on uncommon senses in other WSD datasets
- **Staged Fine-tuning:** train model on two datasets
  - 1st: the **intermediate** training set
  - 2nd: the **target** training set
- Evaluate models on **target** evaluation set

Transfer Learning With FEWS

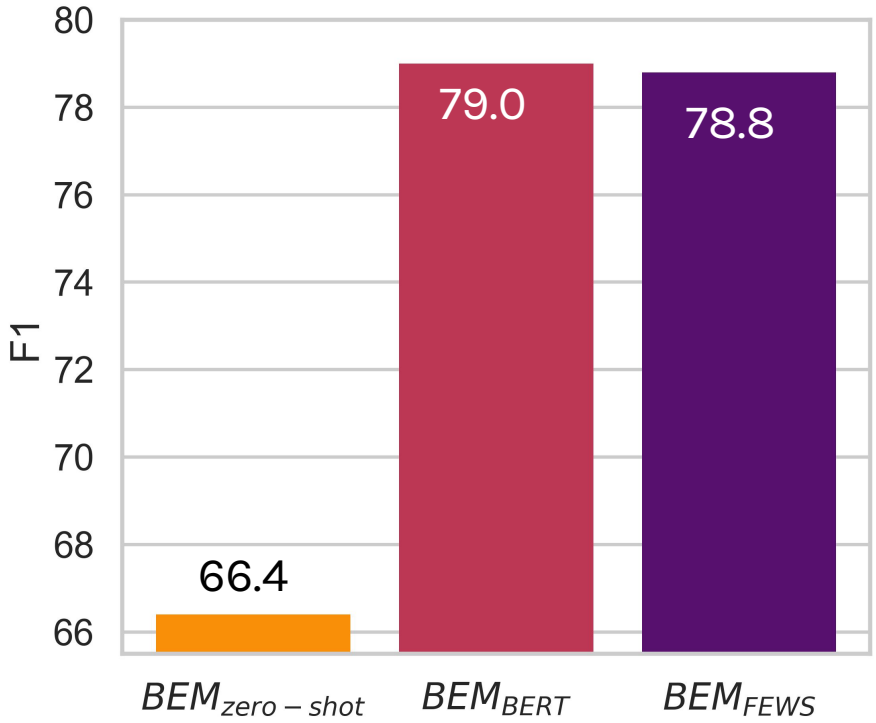
- FEWS -> **intermediate** dataset
- WSD Framework (Raganato et al., 2017) -> **target** dataset

Transfer Learning With FEWS

- FEWS -> **intermediate** dataset
- WSD Framework (Raganato et al., 2017) -> **target** dataset

- Consider performance of biencoder model (**BEM**; Blevins and Zettlemoyer 2020) trained on
  - Only the target dataset (**BEM_{BERT}**)
  - Only the intermediate dataset (**$BEM_{zero\text{-}shot}$**)
  - Both the intermediate and target datasets (**BEM_{FEWS}**)

Transfer Learning With FEWS



WSD Framework Evaluation by Sense Frequency

	MFS	LFS	Zero-shot	
			Words	Senses
WordNet S1	100.0	0.0	84.9	53.9
BEM_{BERT}	**94.1**	52.6	91.2	68.9
BEM_{FEWS}	93.7	52.9	92.2	74.8
$\text{BEM}_{zero-shot}$	72.6	**55.5**	**92.7**	**80.5**

WSD Framework Evaluation by Sense Frequency

	MFS	LFS	Zero-shot	
			Words	Senses
WordNet S1	100.0	0.0	84.9	53.9
BEM_{BERT}	**94.1**	52.6	91.2	68.9
BEM_{FEWS}	93.7	52.9	92.2	74.8
$\text{BEM}_{zero-shot}$	72.6	**55.5**	**92.7**	**80.5**

Takeaways

- **FEWS** is a WSD dataset that provides low-shot training data and evaluation of rare senses.

Takeaways

- **FEWS** is a WSD dataset that provides low-shot training data and evaluation of rare senses.
- All considered baselines lag behind human performance on FEWS, leaving room for future improvement

Takeaways

- **FEWS** is a WSD dataset that provides low-shot training data and evaluation of rare senses.
- All considered baselines lag behind human performance on FEWS, leaving room for future improvement
- Transfer learning experiments demonstrate that FEWS improves performance on uncommon senses in other WSD evaluations.

Takeaways

- **FEWS** is a WSD dataset that provides low-shot training data and evaluation of rare senses.
- All considered baselines lag behind human performance on FEWS, leaving room for future improvement
- Transfer learning experiments demonstrate that FEWS improves performance on uncommon senses in other WSD evaluations.

Questions?

https://www.nlp.cs.washington.edu/fews/

blvns@cs.washington.edu