What do Language Models know about word senses?

Zero-Shot WSD with Language Models and Domain Inventories

Oscar Sainz, Oier Lopez de Lacalle, Eneko Agirre and German Rigau

eman ta zabal zazu

Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

Previous works on WSD using LMs

- SOTA is achieved by fine-tuning LMs on SemCor (Vial et al., GWC 2019).
- Zero-Shot methods are evaluated on lemmas unseen during training, but rely on WSD data to learn the task itself (Lacerra et al., AAAI 2020).
- Present unsupervised and knowledge-based methods do not rely on LMs.

Previous works on WSD using LMs

- What about removing supervised data from LMs?

Previous works on WSD using LMs

- What about removing supervised data from LMs?

	WiC Accuracy
Fine-tuned SOTA	**76.1**
Fine-tuned BERT-Large	69.6
GPT-3 Few-Shot	49.4

(Brown et al., 2020) Language Models are Few-Shot Learners

Previous works on WSD using LMs

- What about removing supervised data from LMs?

	WiC Accuracy
Fine-tuned SOTA	**76.1**
Fine-tuned BERT-Large	69.6
GPT-3 Few-Shot	49.4
Random baseline	50.0

(Brown et al., 2020) [Language Models are Few-Shot Learners]

Recent Advances in DeepLearning for NLP

- 2013 Word Embeddings
- 2018 Transformers & Pretrained Language Models
- 2020 Zero- and Few-Shot prompting
- 2021 Instruction fine-tuning

(Min et al., 2021) Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey

Recent Advances in DeepLearning for NLP

- 2013 Word Embeddings
- 2018 Transformers & Pretrained Language Models
- 2020 Zero- and Few-Shot prompting
- 2021 Instruction fine-tuning

(Min et al., 2021) Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey

__Prompting__ is the practice of adding natural language text, often short phrases, to the input or output to encourage pre-trained models to perform specific tasks.

(Min et al., 2021) Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey

Prompting strategies

EXAMPLE BASED

Translate English to French:

sea otter => loutre de mer

peppermint => menthe poivrée

plush girafe => girafe peluche

cheese => ……………………………………

TEMPLATE BASED

Best pizza ever! It was ………………

great | bad

………… News: OpenAI presents a new model!

World | Sports | Tech

It's snowing. ……………, it's cold.

Yes | Maybe | No

PROXY TASK BASED

premise: I am feeling grouchy.
hypotheses:
It expresses love.
It expresses anger.
It expresses sadness.

C: China has purchased two nuclear submarines from Russia last month.

Q: Who bought something?
A: China
Q: What is bought?
A: Two nuclear submarines.

(Min et al., 2021) Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey

Prompting strategies

EXAMPLE BASED

Translate English to French:

sea otter => loutre de mer

peppermint => menthe poivrée

plush girafe => girafe peluche

cheese => ....

TEMPLATE BASED

Best pizza ever! It was ....

great | bad

.... News: OpenAI presents a new model!

World | Sports | Tech

It's snowing. ...., it's cold.

Yes | Maybe | No

PROXY TASK BASED

premise: I am feeling grouchy.
hypotheses:
It expresses love.
It expresses anger.
It expresses sadness.

C: China has purchased two nuclear submarines from Russia last month.

Q: Who bought something?
A: China
Q: What is bought?
A: Two nuclear submarines.

(Min et al., 2021) Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey

Prompting strategies

EXAMPLE BASED

Translate English to French:

sea otter => loutre de mer

peppermint => menthe poivrée

plush girafe => girafe peluche

cheese => ....

TEMPLATE BASED

Best pizza ever! It was ....

great | bad

.... News: OpenAI presents a new model!

World | Sports | Tech

It's snowing. ...., it's cold.

Yes | Maybe | No

PROXY TASK BASED

premise: I am feeling grouchy.
hypotheses:
It expresses love.
It expresses anger.
It expresses sadness.

C: China has purchased two nuclear submarines from Russia last month.

Q: Who bought something?
A: China
Q: What is bought?
A: Two nuclear submarines.

(Min et al., 2021) Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey

Prompting strategies

Example Based

Translate English to French:

sea otter => loutre de mer

peppermint => menthe poivrée

plush girafe => girafe peluche

cheese => ....

Template Based

Best pizza ever! It was ....

great | bad

.... News: OpenAI presents a new model!

World | Sports | Tech

It's snowing. ...., it's cold.

Yes | Maybe | No

Proxy Task Based

premise: I am feeling grouchy.
hypotheses:
It expresses love.
It expresses anger.
It expresses sadness.

C: China has purchased two nuclear submarines from Russia last month.

Q: Who bought something?
A: China
Q: What is bought?
A: Two nuclear submarines.

(Min et al., 2021) Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey

Prompting strategies

Example Based

Translate English to French:

sea otter => loutre de mer

peppermint => menthe poivrée

plush girafe => girafe peluche

cheese => ....

Template Based

Best pizza ever! It was ....

great | bad

.... News: OpenAI presents a new model!

World | Sports | Tech

It's snowing. ...., it's cold.

Yes | Maybe | No

Proxy Task Based

(Emotion classification as Textual Entailment)

premise: I am feeling grouchy.
hypotheses:
It expresses love.
It expresses anger.
It expresses sadness.

(Argument extraction as Question Answering)

C: China has purchased two nuclear submarines from Russia last month.

Q: Who bought something?
A: China
Q: What is bought?
A: Two nuclear submarines.

(Min et al., 2021) Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey

Textual Entailment as a proxy

PREMISE

Two men on bicycles competing in a race.

(Bowman et al, 2015) A large annotated corpus for learning natural language inference.

Textual Entailment as a proxy

PREMISE

Two men on bicycles competing in a race.

ENTAILMENT HYPOTHESIS

People are riding bikes.

Entailment: the hipothesis is entailed by the premise.

(Bowman et al, 2015) A large annotated corpus for learning natural language inference.

Textual Entailment as a proxy

PREMISE

Two men on bicycles competing in a race.

ENTAILMENT HYPOTHESIS

People are riding bikes.

Entailment: the hypothesis is entailed by the premise.

NEUTRAL HYPOTHESIS

Men are riding bicycles on the street.

Neutral: the hypothesis can not be entailed by the premise.

(Bowman et al, 2015) A large annotated corpus for learning natural language inference.

Textual Entailment as a proxy

PREMISE

Two men on bicycles competing in a race.

ENTAILMENT HYPOTHESIS

People are riding bikes.

NEUTRAL HYPOTHESIS

Men are riding bicycles on the street.

CONTRADICTION HYPOTHESIS

Few people are catching fish.

Entailment: the hipothesis is entailed by the premise.

Neutral: the hipothesis can not be entailed by the premise.

Contradiction: the hipothesis contradicts the premise.

(Bowman et al, 2015) A large annotated corpus for learning natural language inference.

Domain Labelling with Textual Entailment

hospital: a health facility where patients receive threatment.

BIOLOGY	BUSINESS	CULTURE	ECONOMY	LEGAL	MEDICINE	POLITICS

(Yin et al, 2019) Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach
(Sainz and Rigau, 2021) Ask2Transformers: Zero-Shot domain labelling with Pretrained Language Models.

Domain Labelling with Textual Entailment

PREMISE

hospital: a health facility where patients receive threatment.

BIOLOGY | BUSINESS | CULTURE | ECONOMY | LEGAL | MEDICINE | POLITICS

(Yin et al, 2019) Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach
(Sainz and Rigau, 2021) Ask2Transformers: Zero-Shot domain labelling with Pretrained Language Models.

Domain Labelling with Textual Entailment

PREMISE

hospital: a health facility where patients receive threatment.

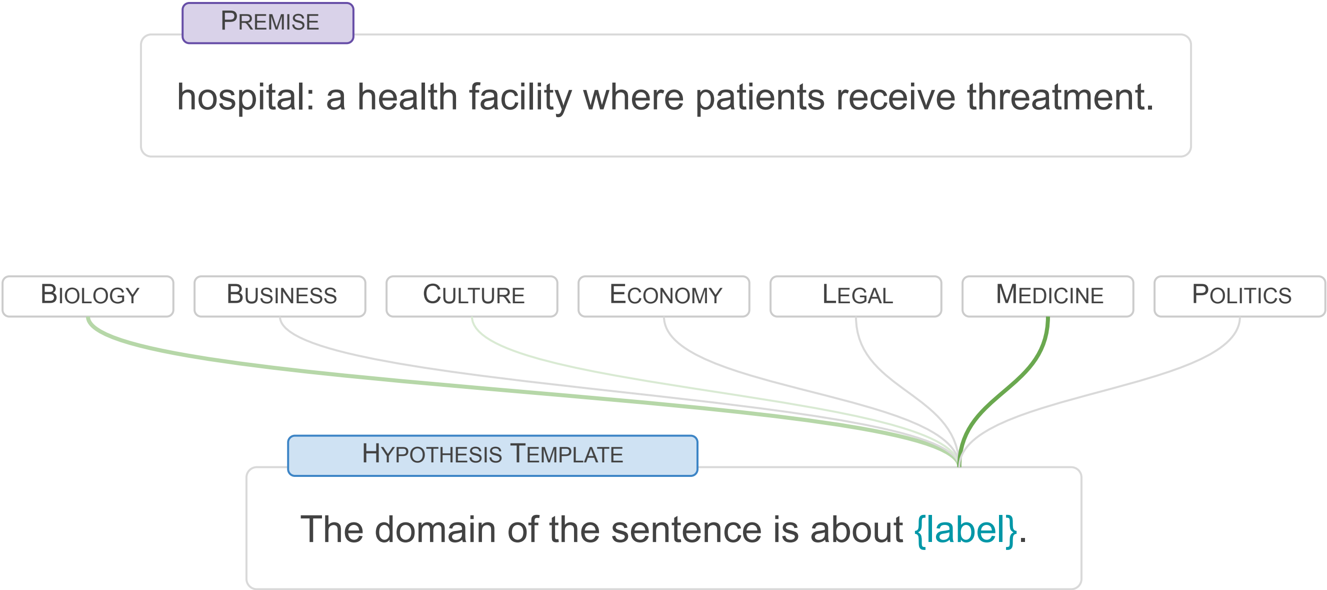
BIOLOGY	BUSINESS	CULTURE	ECONOMY	LEGAL	MEDICINE	POLITICS

HYPOTHESIS TEMPLATE

The domain of the sentence is about {label}.

(Yin et al, 2019) Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach
(Sainz and Rigau, 2021) Ask2Transformers: Zero-Shot domain labelling with Pretrained Language Models.

Domain Labelling with Textual Entailment



(Yin et al, 2019) Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach
(Sainz and Rigau, 2021) Ask2Transformers: Zero-Shot domain labelling with Pretrained Language Models.

The task of Word Sense Disambiguation

The medicine can only be obtained with a prescription.

The task of Word Sense Disambiguation

The medicine can only be obtained with a prescription.

6788565-n

6788565-n: directions prescribed beforehand; the action of prescribing authoritative rules or directions.

The task of Word Sense Disambiguation

The medicine can only be obtained with a prescription.

6788565-n

3999280-n

3999280-n: a drug that is available only with written instructions from a doctor or dentist to a pharmacy.

The task of Word Sense Disambiguation

The medicine can only be obtained with a prescription.

6788565-n

3999280-n

6366002-n

6366002-n: written instructions for an optician on the lenses for a given person.

The task of Word Sense Disambiguation

The medicine can only be obtained with a prescription.

6788565-n

3999280-n

6366002-n

6365808-n

6365808-n: written instructions from a physician or dentist to a druggist concerning the form and dosage of a drug to be issued to a given patient.

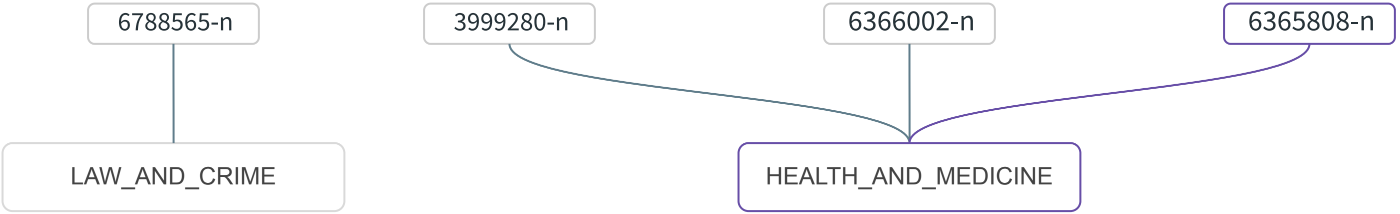
The task of Word Sense Disambiguation

The medicine can only be obtained with a prescription.

(Lacerra et al, 2020) CSI: A Coarse Sense Inventory for 85% Word Sense Disambiguation

The task of Word Sense Disambiguation

The medicine can only be obtained with a prescription.



(Lacerra et al, 2020) CSI: A Coarse Sense Inventory for 85% Word Sense Disambiguation

Domain Inventories

BabelDomains

- Unified domain information for Wikipedia, WordNet and BabelNet.
- Inherits from Wikipedia domains.
- 34 **coarse** domain labels.
- Semi-automatically annotated.

Coarse Sense Inventory

- Created to reduce the granularity of WordNet synsets.
- High agreement among annotators.
- 45 domain labels.
- Manually annotated.

WordNet Domains

- Hierarchical domain definition.
- Domain information for WordNet synsets.
- 160 **fine-grained** domain labels.
- Due to the high granularity and hierarchical nature we kept only 60 labels.

Domain Inventories

BabelDomains

- Unified domain information for Wikipedia, WordNet and BabelNet.
- Inherits from Wikipedia domains.
- 34 **coarse** domain labels.
- Semi-automatically annotated.

Coarse Sense Inventory

- Created to reduce the granularity of WordNet synsets.
- High agreement among annotators.
- 45 domain labels.
- Manually annotated.

WordNet Domains

- Hierarchical domain definition.
- Domain information for WordNet synsets.
- 160 **fine-grained** domain labels.
- Due to the high granularity and hierarchical nature we kept only 60 labels.

Domain Inventories

BabelDomains

- Unified domain information for Wikipedia, WordNet and BabelNet.
- Inherits from Wikipedia domains.
- 34 **coarse** domain labels.
- Semi-automatically annotated.

Coarse Sense Inventory

- Created to reduce the granularity of WordNet synsets.
- High agreement among annotators.
- 45 domain labels.
- Manually annotated.

WordNet Domains

- Hierarchical domain definition.
- Domain information for WordNet synsets.
- 160 **fine-grained** domain labels.
- Due to the high granularity and hierarchical nature we kept only 60 labels.

Domain Inventories

BabelDomains

- Unified domain information for Wikipedia, WordNet and BabelNet.
- Inherits from Wikipedia domains.
- 34 **coarse** domain labels.
- Semi-automatically annotated.

Coarse Sense Inventory

- Created to reduce the granularity of WordNet synsets.
- High agreement among annotators.
- 45 domain labels.
- Manually annotated.

WordNet Domains

- Hierarchical domain definition.
- Domain information for WordNet synsets.
- 160 **fine-grained** domain labels.
- Due to the high granularity and hierarchical nature we kept only 60 labels.

Word Sense Disambiguation with Textual Entailment

The medicine can only be obtained with a prescription.

prescription

Target word   Ambiguous words

Word Sense Disambiguation with Textual Entailment

6788565-n

3999280-n

6366002-n

6365808-n

prescription

The medicine can only be obtained with a prescription.

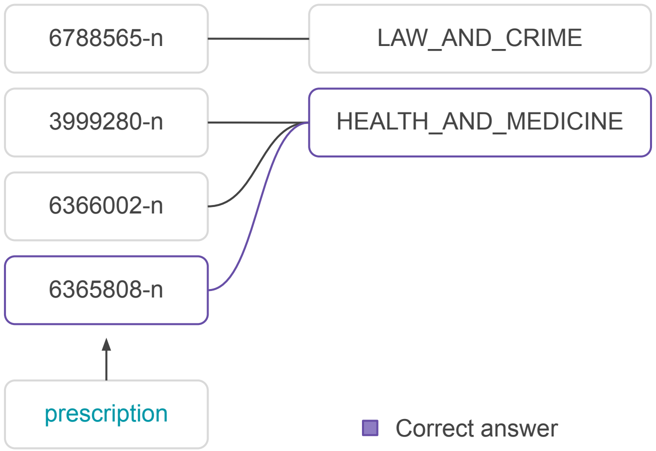
Correct answer

Target word

Ambiguous words

Word Sense Disambiguation with Textual Entailment

Label simplification using Domain Inventories:

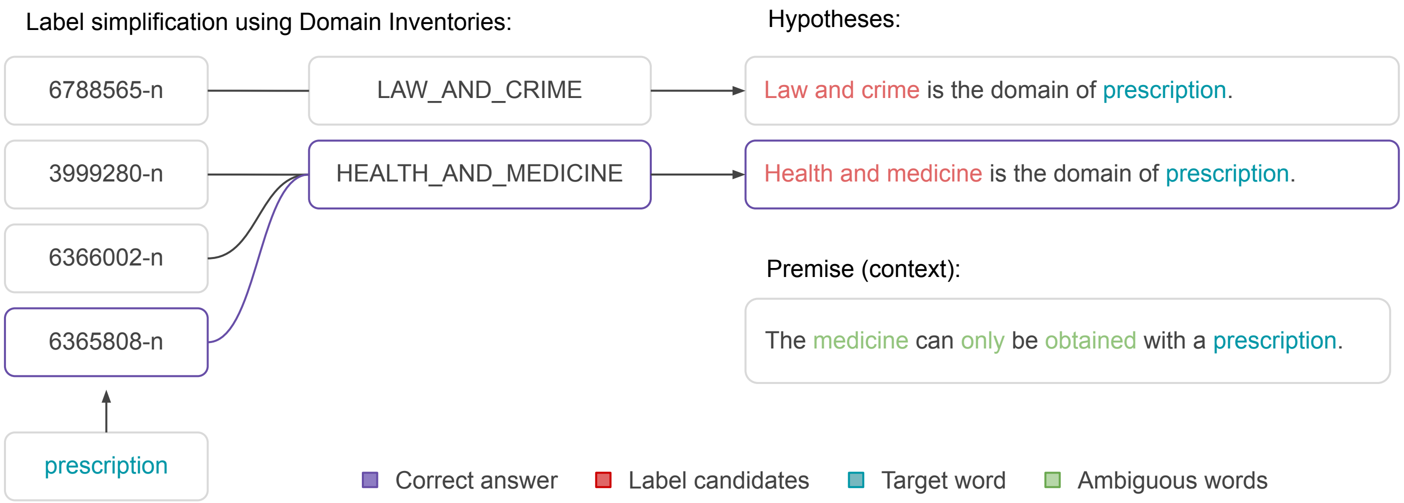


Correct answer

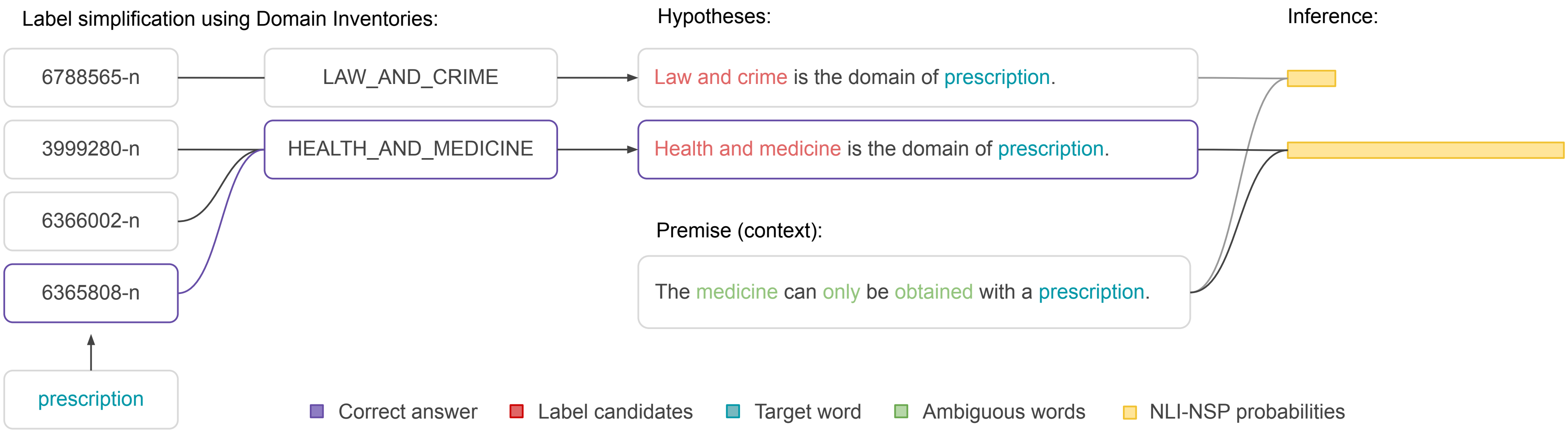
The medicine can only be obtained with a prescription.

Target word   Ambiguous words

Word Sense Disambiguation with Textual Entailment



Word Sense Disambiguation with Textual Entailment



Experimental setup

- 2 different models:

Experimental setup

- 2 different models:
  - BERT: a pretrained Masked Language Model along with the Next Sentence Prediction objective.

Experimental setup

- 2 different models:
  - BERT: a pretrained Masked Language Model along with the Next Sentence Prediction objective.
  - RoBERTa: a pretrained Masked Language Model (similar to BERT) but for larger number of steps.

Experimental setup

- 2 different models:
  - BERT and RoBERTa
- 2 different fine-tuning tasks:
  - NSP: Next Sentence Prediction is the task of predicting whether a sentence follows another or not.

Experimental setup

- 2 different models:
  - BERT and RoBERTa
- 2 different fine-tuning tasks:
  - NSP: Next Sentence Prediction is the task of predicting whether a sentence follows another or not.
  - Textual Entailment: the task of predicting the entailment relation between premises and hypotheses, also known as NLI.

Experimental setup

- 2 different models:
    - BERT and RoBERTa
- 2 different training objectives:
    - Next Sentence Prediction (NSP) and Textual Entailment (NLI)
- Different pre-training data regimes for Textual Entailment:

Experimental setup

- 2 different models:
  - BERT and RoBERTa
- 2 different training objectives:
  - Next Sentence Prediction (NSP) and Textual Entailment (NLI)
- Different pre-training data regimes for Textual Entailment:
  - Using just MNLI dataset.

Experimental setup

- 2 different models:
  - BERT and RoBERTa
- 2 different training objectives:
  - Next Sentence Prediction (NSP) and Textual Entailment (NLI)
- Different pre-training data regimes for Textual Entailment:
  - Using just MNLI dataset.
  - Using several Textual Entailment datasets: SNLI, MNLI, Fever-NLI and aNLI.

Experimental setup

- 2 different models:
  - BERT and RoBERTa
- 2 different training objectives:
  - Next Sentence Prediction (NSP) and Textual Entailment (NLI)
- Different pre-training data regimes for Textual Entailment:
  - NLI: just MNLI
  - NLI*: MNLI, SNLI, Fever-NLI and ANLI

Are language models able to discriminate domains in sense glosses?

PROMPT

{gloss} | The domain of the sentence is about {label}

Are language models able to discriminate domains in sense glosses?

PROMPT

{gloss} | The domain of the sentence is about {label}

cell: (biology) the basic structural and functional unit of all organisms;

Are language models able to discriminate domains in sense glosses?

PROMPT

{gloss} | The domain of the sentence is about {label}

hint!

cell: (biology) the basic structural and functional unit of all organisms;

Are language models able to discriminate domains in sense glosses?

PROMPT

{gloss} | The domain of the sentence is about {label}

hint!

cell: ~~(biology)~~ the basic structural and functional unit of all organisms;

Are language models able to discriminate domains in sense glosses?

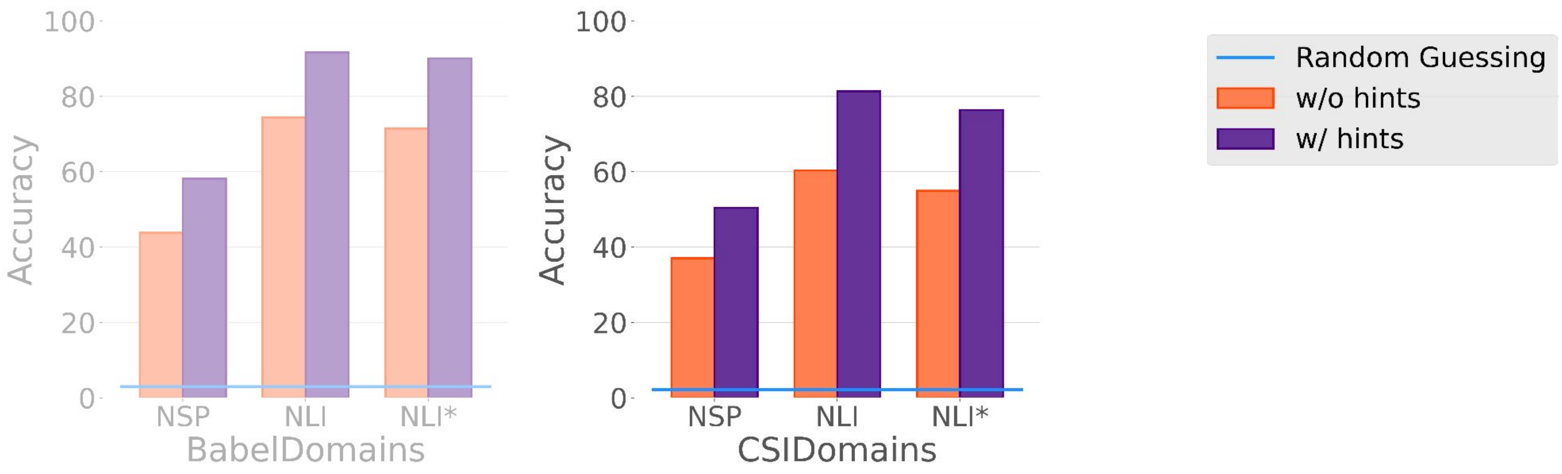
PROMPT

{gloss} | The domain of the sentence is about {label}

Are language models able to discriminate domains in sense glosses?

PROMPT

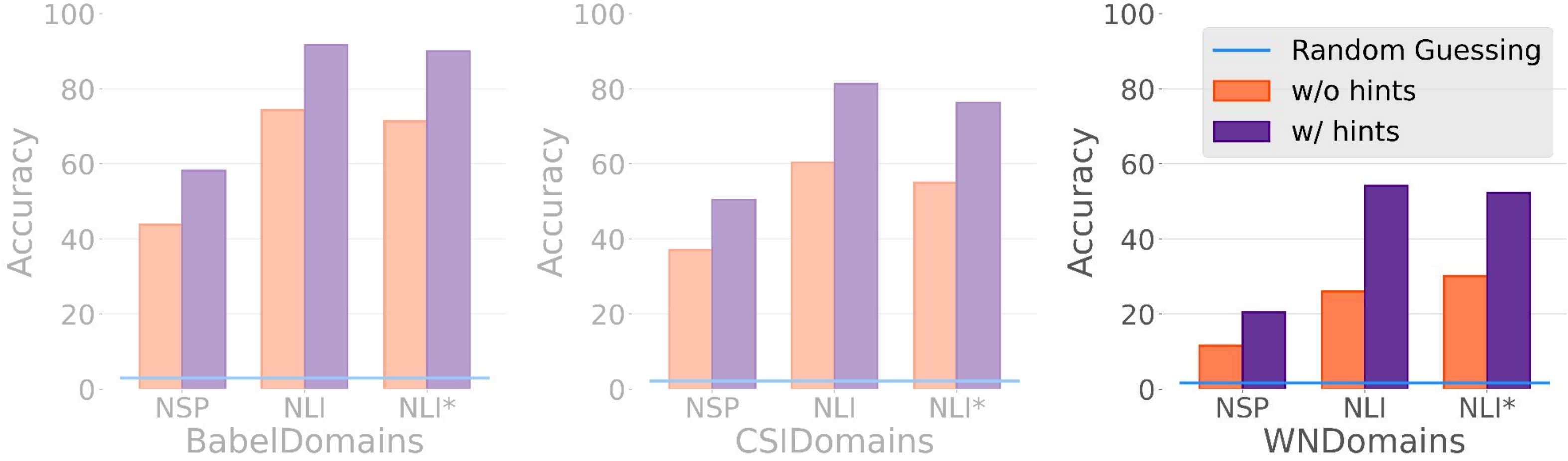
{gloss} | The domain of the sentence is about {label}



Are language models able to discriminate domains in sense glosses?

PROMPT

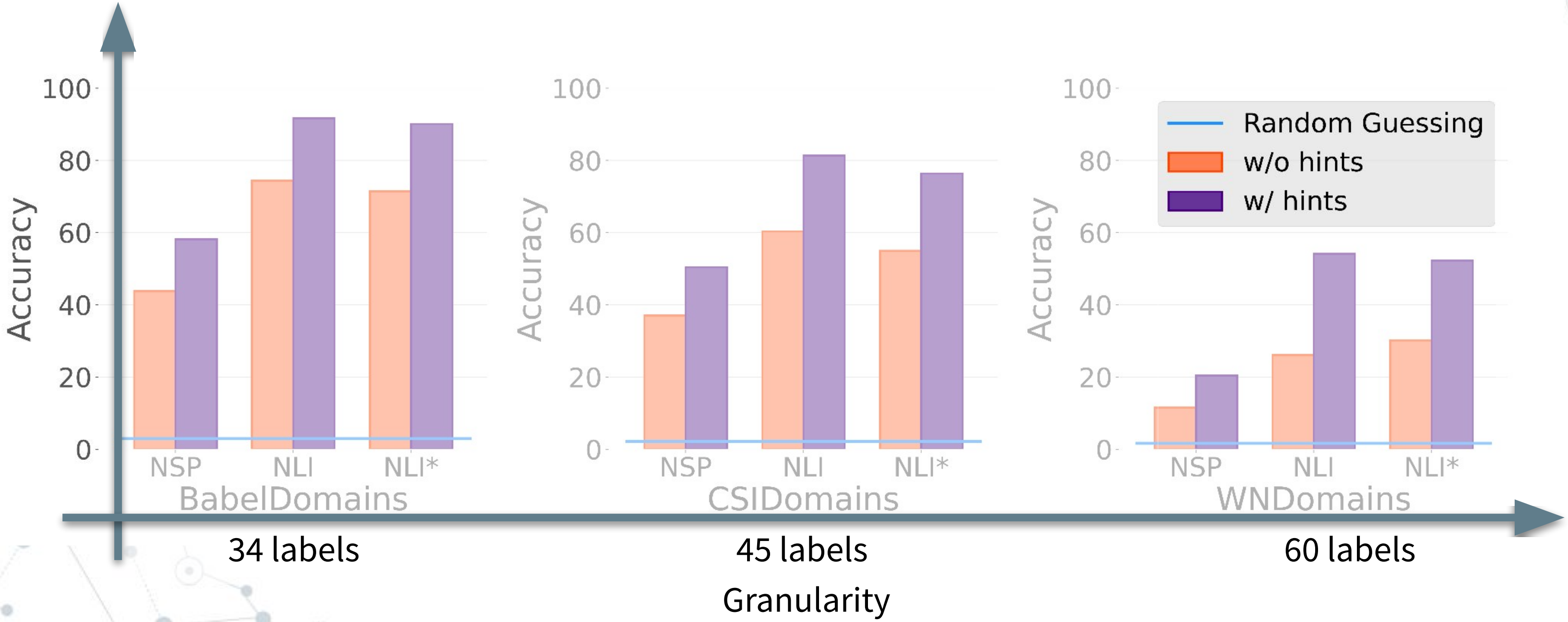
{gloss} | The domain of the sentence is about {label}



Are language models able to discriminate domains in sense glosses?

PROMPT

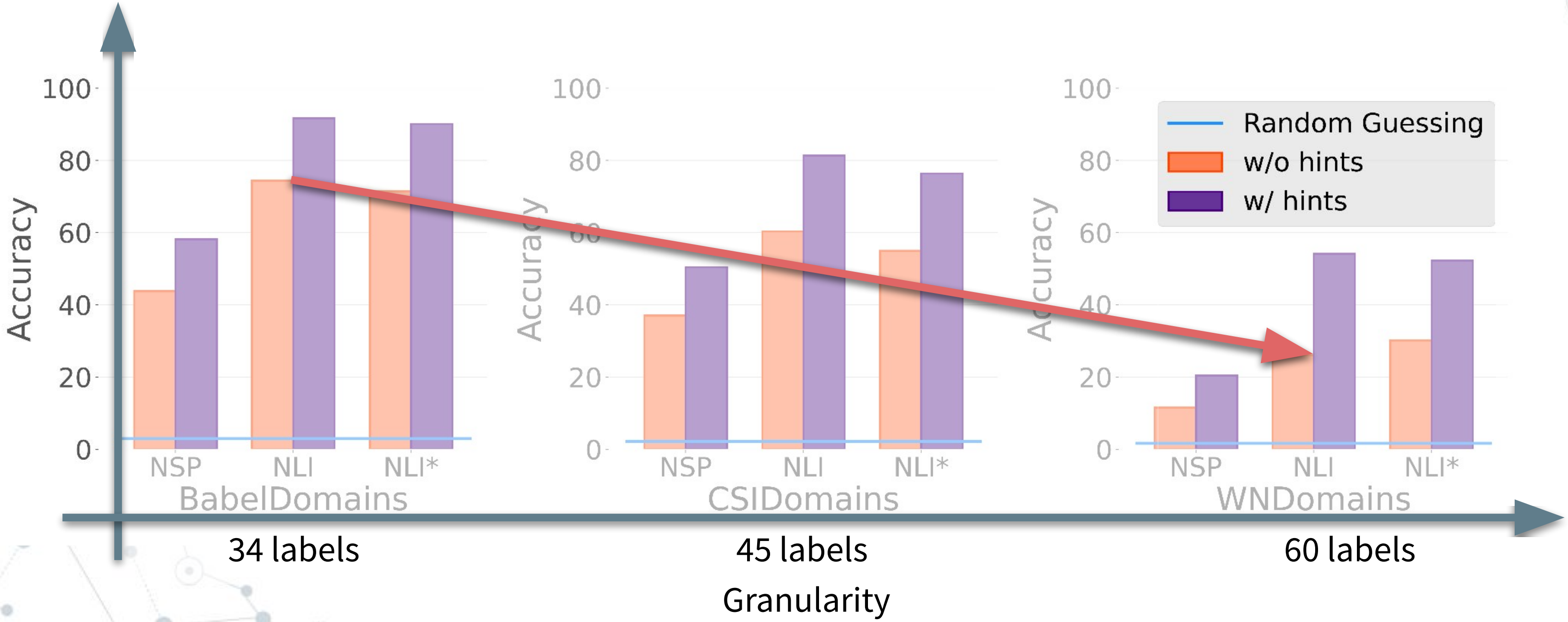
{gloss} | The domain of the sentence is about {label}



Are language models able to discriminate domains in sense glosses?

PROMPT

{gloss} | The domain of the sentence is about {label}



Do Language Models know about Word Senses?

SENT PROMPT

{context} | The domain of the sentence is about {label}.

WORD PROMPT

{context} | {label} is the domain of {word}.

Do Language Models know about Word Senses?

SENT PROMPT

{context} | The domain of the sentence is about {label}.

WORD PROMPT

{context} | {label} is the domain of {word}.

Supervised — Random Guessing — sent prompt — word prompt

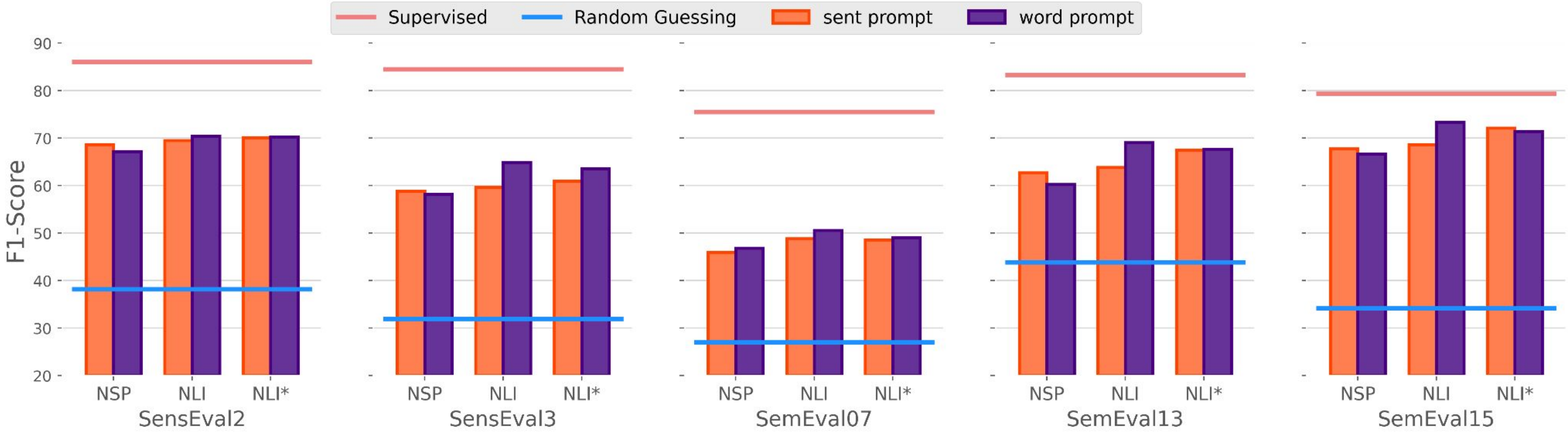
Do Language Models know about Word Senses?

SENT PROMPT

{context} | The domain of the sentence is about {label}.

WORD PROMPT

{context} | {label} is the domain of {word}.



Do Language Models perform differently depending on the word category?

SENT PROMPT

{context} | The domain of the sentence is about {label}.

WORD PROMPT

{context} | {label} is the domain of {word}.

Model	Noun	Adj	Verb	Adv	All
Random	40.7	48.4	23.7	59.1	38.8
Sentence prompt					
NSP	60.3	84.9	50.4	86.6	62.6
NLI	64.3	86.2	54.8	86.4	66.1
NLI*	65.0	85.9	55.0	85.3	66.4
Word prompt					
NSP	59.4	84.8	50.2	86.4	61.9
NLI	**66.2**	**86.8**	**57.0**	**87.3**	**67.8**
NLI*	65.3	85.5	55.7	85.5	66.8

Table 3: F1-Scores per word category

Do Language Models perform differently depending on the word category?

SENT PROMPT

{context} | The domain of the sentence is about {label}.

WORD PROMPT

{context} | {label} is the domain of {word}.

Model	Noun	Adj	Verb	Adv	All
Random	40.7	48.4	23.7	59.1	38.8
Sentence prompt					
NSP	60.3	84.9	50.4	86.6	62.6
NLI	64.3	86.2	54.8	86.4	66.1
NLI*	65.0	85.9	55.0	85.3	66.4
Word prompt					
NSP	59.4	84.8	50.2	86.4	61.9
NLI	**66.2**	**86.8**	**57.0**	**87.3**	**67.8**
NLI*	65.3	85.5	55.7	85.5	66.8

Table 3: F1-Scores per word category

Do Language Models perform differently depending on the word category?

SENT PROMPT

{context} | The domain of the sentence is about {label}.

WORD PROMPT

{context} | {label} is the domain of {word}.

Model	Noun	Adj	Verb	Adv	All
Random	40.7	48.4	23.7	59.1	38.8
Sentence prompt					
NSP	60.3	84.9	50.4	86.6	62.6
NLI	64.3	86.2	54.8	86.4	66.1
NLI*	65.0	85.9	55.0	85.3	66.4
Word prompt					
NSP	59.4	84.8	50.2	86.4	61.9
NLI	**66.2**	**86.8**	**57.0**	**87.3**	**67.8**
NLI*	65.3	85.5	55.7	85.5	66.8

Table 3: F1-Scores per word category

Do Language Models perform differently depending on the word category?

SENT PROMPT

{context} | The domain of the sentence is about {label}.

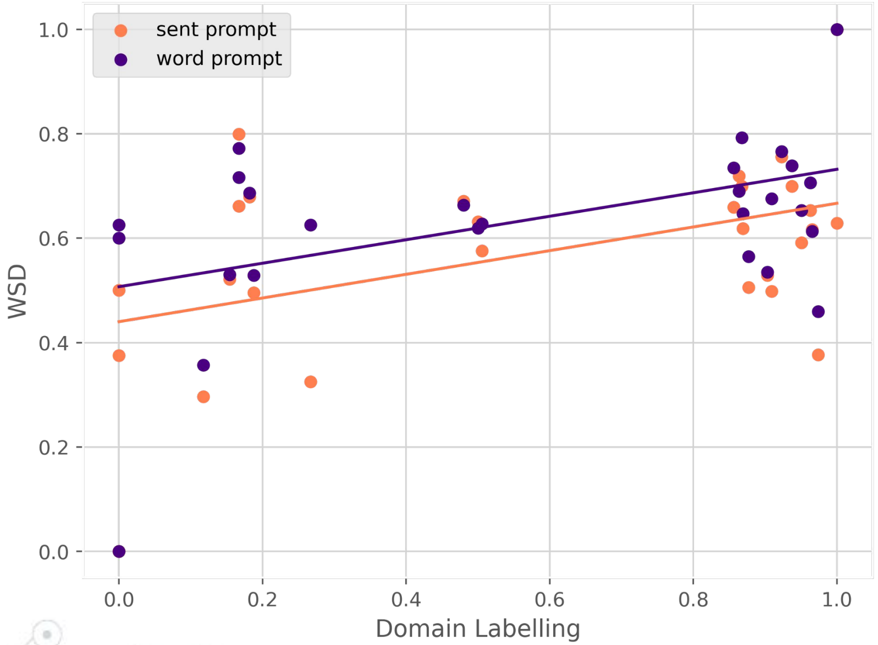
WORD PROMPT

{context} | {label} is the domain of {word}.

Model	Noun	Adj	Verb	Adv	All
Random	40.7	48.4	23.7	59.1	38.8
Sentence prompt					
NSP	60.3	84.9	50.4	86.6	62.6
NLI	64.3	86.2	54.8	86.4	66.1
NLI*	65.0	85.9	55.0	85.3	66.4
Word prompt					
NSP	59.4	84.8	50.2	86.4	61.9
NLI	**66.2**	**86.8**	**57.0**	**87.3**	**67.8**
NLI*	65.3	85.5	55.7	85.5	66.8

Table 3: F1-Scores per word category

To what extent does the performance on Domain Labelling affects WSD?



	Dom Lab.	WSD_{sent}	WSD_{word}
Dom Lab.	1.00	0.32	0.41
WSD_{sent}	0.32	1.00	0.81
WSD_{word}	0.41	0.81	1.00

Table 4: Spearman's correlation of F1-Scores between tasks using shared labels. The scores correspond to the NLI model.

Conclusions

- We present an approach for Zero-Shot WSD using LMs.

Conclusions

- We present an approach for Zero-Shot WSD using LMs.
- We showed that LMs have some notion of senses even without training them for that.

Conclusions

- We present an approach for Zero-Shot WSD using LMs.
- We showed that LMs have some notion of senses even without training them for that.
- We show that errors from Domain Labelling are not directly propagated to WSD.
-

Conclusions

- We present an approach for Zero-Shot WSD using LMs.
- We showed that LMs have some notion of senses even without training them for that.
- We show that errors from Domain Labelling are not directly propagated to WSD.
- For the future, would be interesting to analyze more recent (and bigger) LMs.

Conclusions

- We present an approach for Zero-Shot WSD using LMs.
- We showed that LMs have some notion of senses even without training them for that.
- We show that errors from Domain Labelling are not directly propagated to WSD.
- For the future, would be interesting to analyze more recent (and bigger) LMs.

What do Language Models know about word senses?

Zero-Shot WSD with Language Models and Domain Inventories

Oscar Sainz, Oier Lopez de Lacalle, Eneko Agirre and German Rigau

eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea

HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology