Ambiguity Meets Uncertainty: Investigating Uncertainty Estimation for Word Sense Disambiguation

Zhu Liu, Ying Liu

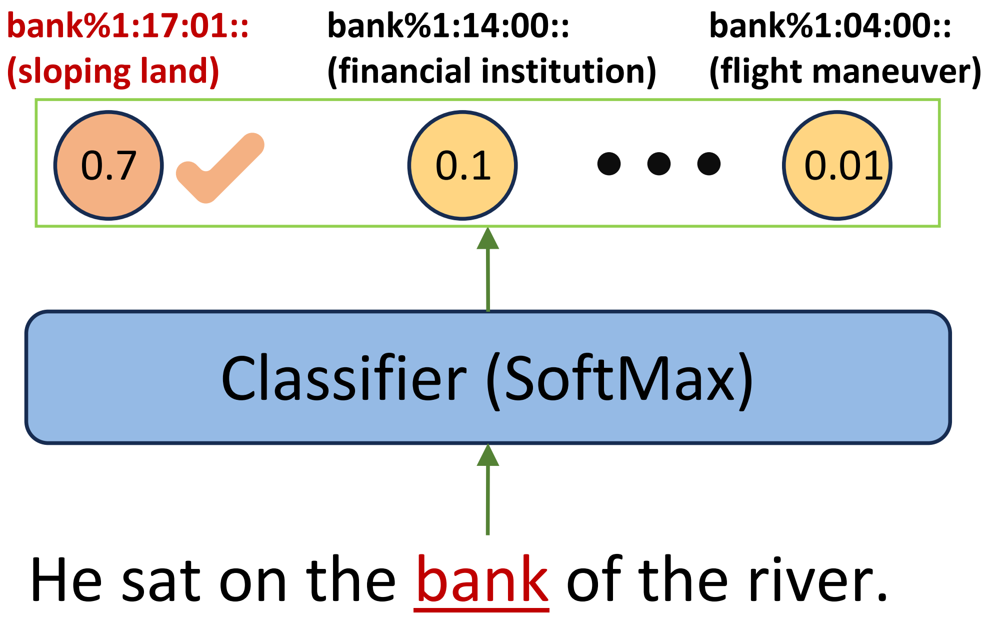
liuzhu22@mails.tsinghua.edu.cn

yingliu@tsinghua.edu.cn

Introduction

Task and Problem

- A deterministic classification task for Word sense disambiguation (WSD).



Introduction

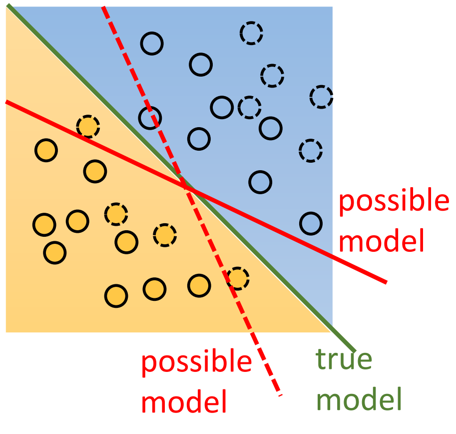
Task and Problem

- A deterministic classification task for Word sense disambiguation (WSD).
- Probability score after SoftMax is poorly calibrated
- Fail to estimate uncertainty

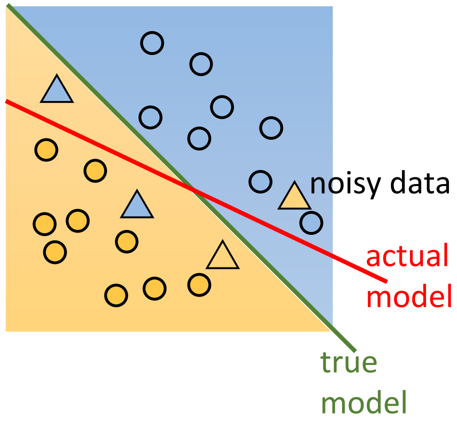
Introduction

Task and Problem

- A deterministic classification task for Word
- Probability score after Softmax is not well-
- Fail to estimate uncertainty



(a) Model uncertainty



(b) Data uncertainty

- Model uncertainty: varied models due to inadequate data
- Data uncertainty: random results due to inherent noise

Introduction

Ambiguity meets Uncertainty

- WSD requires uncertainty estimation
- Model uncertainty

  Imbalanced sense distribution (Most-Frequent-sense bias)

  Domain shift (Different genres, language styles…)
- Data uncertainty

  Imperfect annotations with relatively low agreement (~80%)

  Literal vs. non-literal understandings

Introduction

Contributions

- To compare the conventional probability of the model output with the other three uncertainty scores
- To design test scenarios to evaluate model and data uncertainty
- To analyze which lexical properties affect uncertainty estimation.

Evaluation

Uncertainty Scores

- Model: a SOTA WSD model (MLS [Conia and Navigli, 2021])
- Test Datasets: the Unified Evaluation Framework for English all-words(Senseval-2, Senseval-3,SemEval-2007, SemEval-2013, and SemEval-2015)
- UE scores: MP, SMP, PV and BALD

  MP: negative Softmax output; Other scores: MC Dropout Sample statistics
- Metrics: RCC (risk courage curve) and RPP (reversed pair proportion)

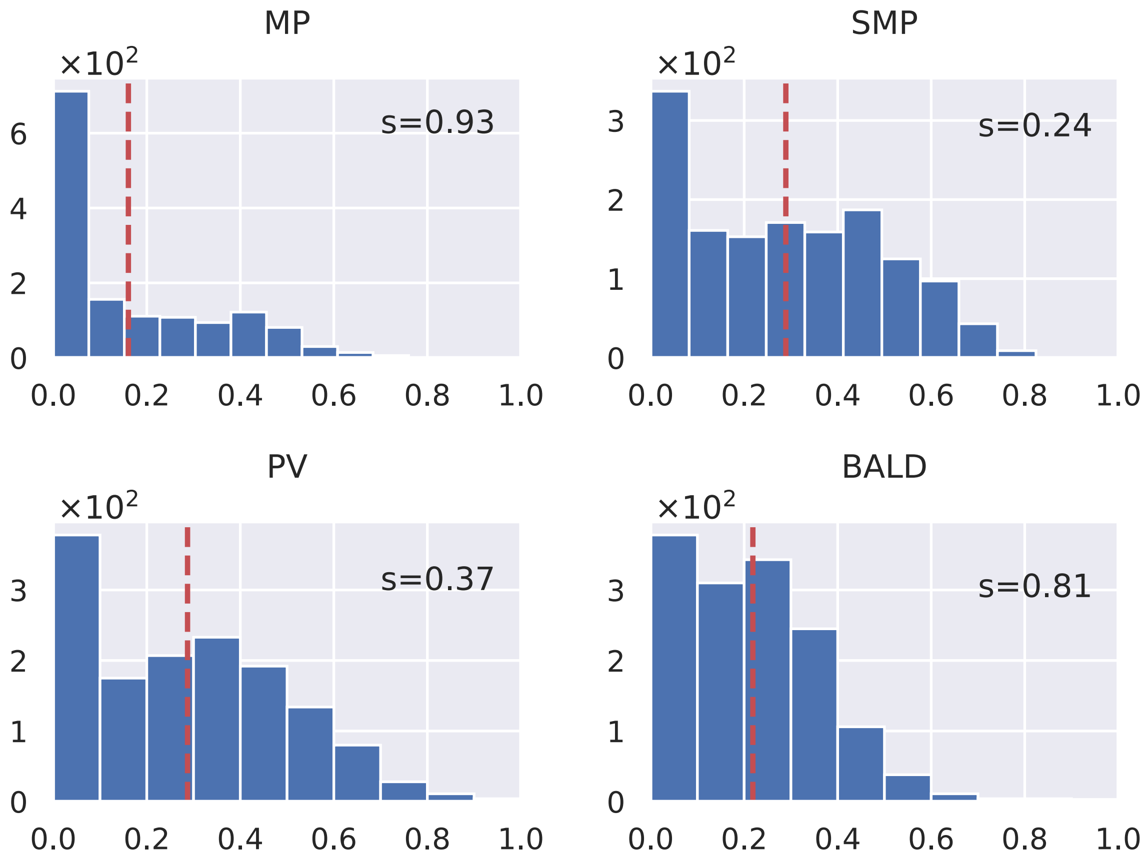
  RCC: cumulative misclassifications according to uncertainty levels

  RPP: Disagreement samples between uncertainty and loss values

Evaluation

Uncertainty Scores

- Question: which UE score is better?
- The distribution of four UE scores on misclassified instances of all datasets.
- Sample-based score SMP better than MP with a more balanced distribution
- MP tends to be over-confident



Evaluation

Uncertainty Scores

UE Score	Senseval-2		Senseval-3		SemEval-07		SemEval-13		SemEval-15	
	RCC ↓	RPP ↓	RCC ↓	RPP ↓	RCC ↓	RPP ↓	RCC ↓	RPP ↓	RCC ↓	RPP ↓
MP	**5.69**	9.50	7.11	10.37	**8.68**	11.40	5.78	8.02	**5.02**	**11.07**
SMP	5.78	**9.14**	**7.10**	**9.83**	8.81	**10.83**	**5.59**	**7.88**	5.34	11.16
PV	6.11	11.47	7.50	12.40	9.93	16.00	5.97	10.22	5.62	13.11
BALD	6.00	11.09	7.46	11.99	9.36	14.73	5.83	10.02	5.48	12.77

Table 1: UE score comparisons on five standard WSD datasets.

UE Score	NOUN		VERB		ADJ		ADV		ALL	
	RCC ↓	RPP ↓	RCC ↓	RPP ↓	RCC ↓	RPP ↓	RCC ↓	RPP ↓	RCC ↓	RPP ↓
MP	6.06	**7.47**	14.08	18.20	5.15	**8.25**	3.70	4.89	6.13	9.78
SMP	**4.94**	7.66	**13.76**	**17.45**	**4.39**	8.35	**2.65**	**4.85**	**6.11**	**9.44**
PV	6.25	9.17	15.38	22.02	4.97	9.37	3.20	5.33	6.48	11.91
BALD	5.18	9.39	14.42	20.96	4.59	9.80	2.66	5.56	6.36	11.52

Table 2: UE score comparisons on all the datasets with different kinds of POS.

- SMP has an advantage over other scores.

Evaluation

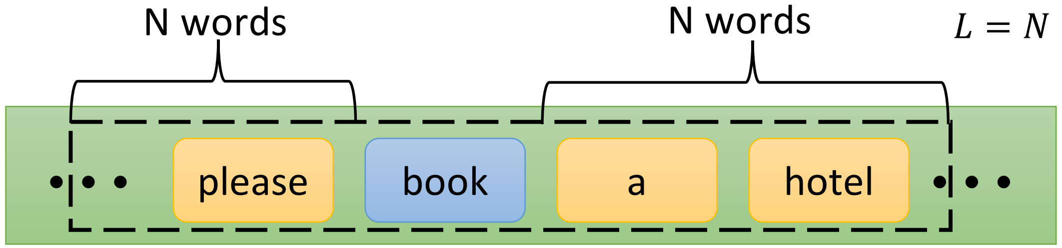
Data Uncertainty

- Controllable context to simulate partial observations
- Window-controlled context

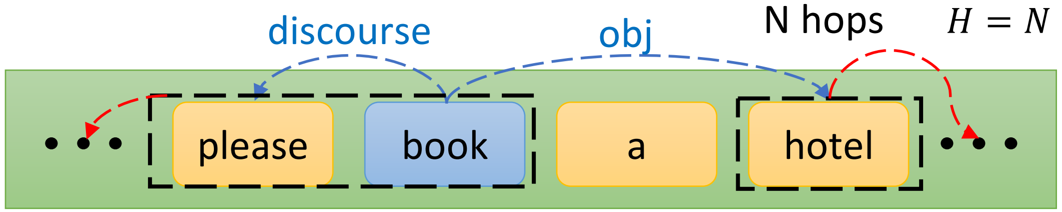
  N linear neighboring words
- Syntax-controlled context

  hierarchical neighboring words connected by universal dependency

  N hops



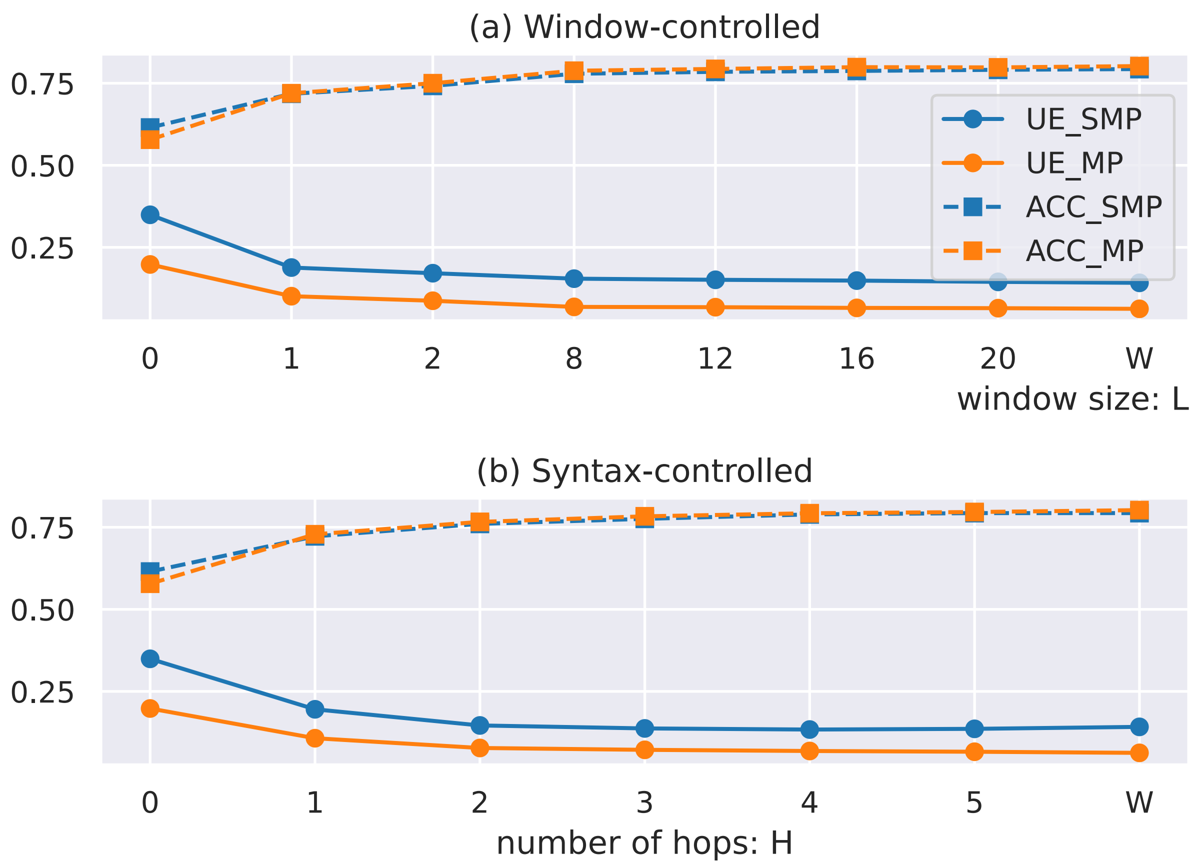
(a) window-controlled context



(b) syntax-controlled context

Evaluation
Data Uncertainty

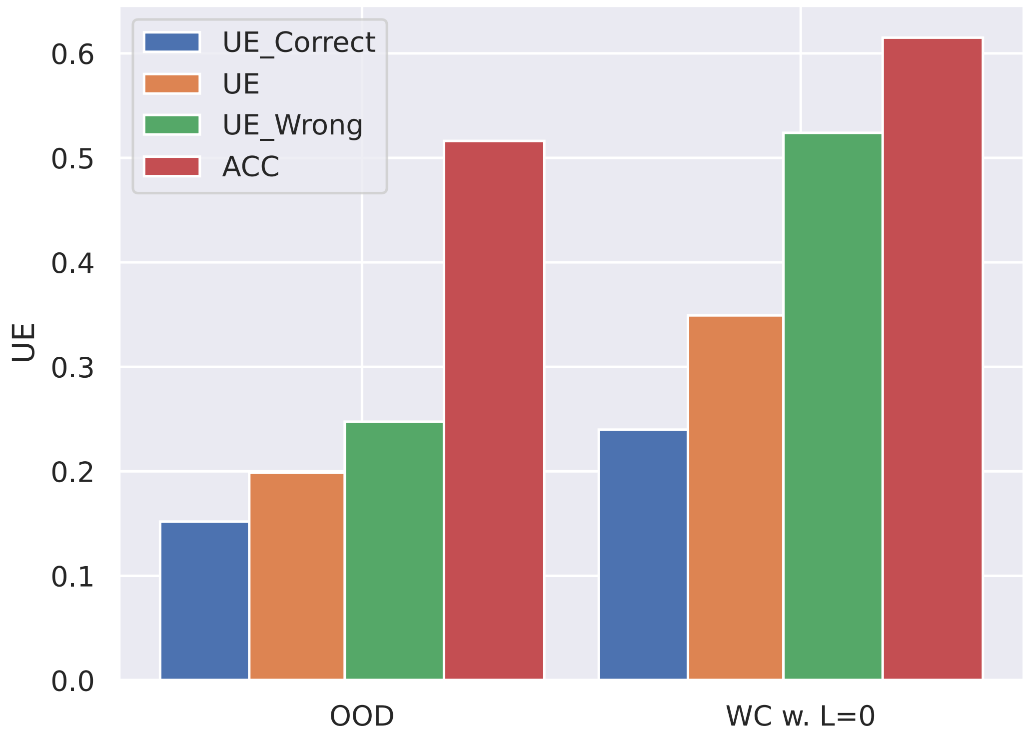
- **How does the model capture DU?**
- We expect that with the larger window size or number of hops, the more accurate and the more uncertain the model will be.
- SMP captures data uncertainty better



Evaluation

Model Uncertainty

- How does the model capture MU?
- Out-of-distributed dataset: 42D [Maru et al., 2022]
- Lower uncertainty than the most (data) uncertain case
- SMP underestimates model uncertainty



Uncertainty and accuracy (F1) scores for model uncertainty (OOD) and data uncertainty (without any context) scenarios.

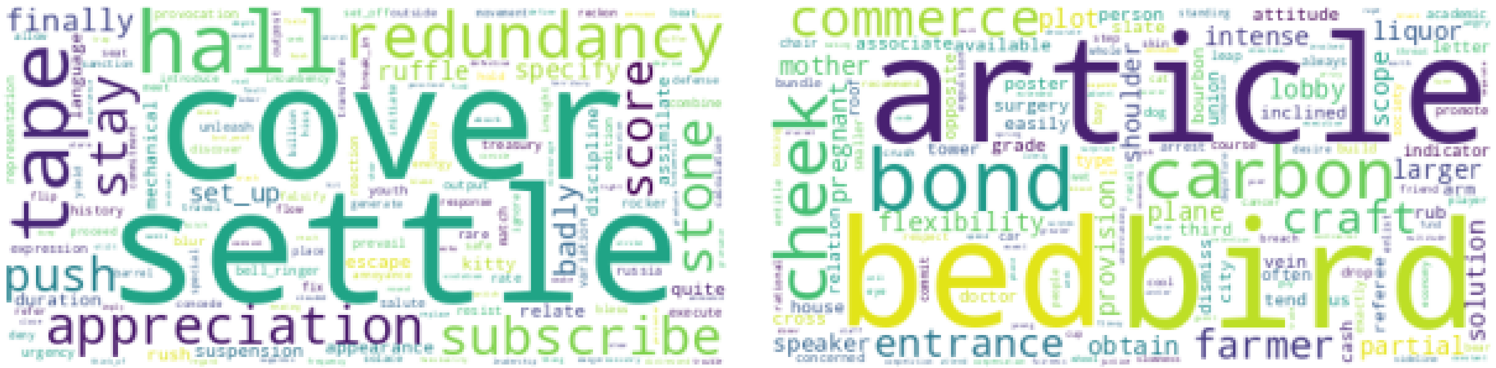
Qualitative Results

- Words with different levels of uncertainty
- Most uncertain words, e.g., *settle, cover*

  Most certain words, e.g., *article, bed, bird*
- **Which lexical properties affect uncertainty estimation?**

(a) Most uncertain lemmas

(b) Most certain lemmas



Analysis

Effects on Uncertainty

- Syntactic Category
- Morphology
- Sense Granularity
- Semantic relation

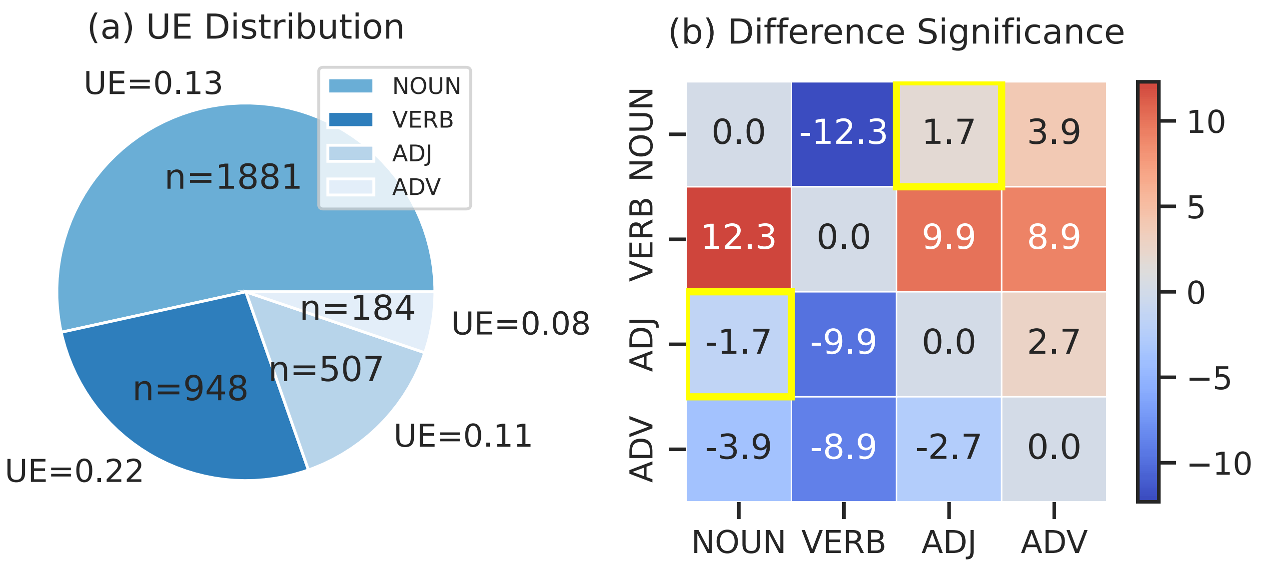
Question: Given different word groups split by the uncertainty level, is there significant difference in their mean values between each other?

- N splits for different effects, considering the trade-off of level granularity and sample sparsity
- T-test with p-value of 5%

Analysis

Effects on Uncertainty

- Syntactic Category
- Morphology
- Sense Granularity
- Semantic relation



Significant difference among different syntactic categories

Except for the NOUN-ADJ pair, verbal instances are more significantly uncertain than NOUN or ADJ, while ADV has the least uncertainty.

Analysis

Effects on Uncertainty

- Syntactic Category
- Morphology ✓

  number of morphemes (nMorph)
- Sense Granularity ✓

Number of ground-truth senses (nGT)

Number of candidate senses (nPD)

- Semantic relation ✓

Hyponymy for nouns (dHypo)

Synonym (dSyno) ✗

Effect	Condition	Agg.	Uncertainty Estimation L1	L2	L3	Difference Significance L1 ↔ L2	L1 ↔ L3	L2 ↔ L3
nMorph	nGT=1, POS=NOUN	L	0.13	0.11	0.07	**1.44e-2**	**1.35e-8**	**5e-4**
	nGT=1, POS=VERB		0.22	0.19	0.13	7.61e-2	**6.04e-4**	6.6e-2
	nGT=1, POS=ADJ		0.11	0.08	0.10	**3.6e-2**	4.21e-1	4.40e-1
	nGT=1, POS=ADV		0.11	0.06	0.02	7.6e-2	**6.04e-4**	6.60e-2
nGT	-	I	0.12	0.22	-	**1.61e-22**	-	-
nPD	nGT=1	L	0.04	0.16	0.22	**6.22e-96**	**3.42e-135**	**5.01e-10**
dHypo	nGT=1, POS=NOUN	L	0.14	0.12	0.09	**1.43e-2**	**1.91e-6**	**6e-3**
dSyno	nGT=1	S	0.14	0.14	0.14	5.55	5.38	5.67

Significant difference among different levels in terms of various effects

Conclusion

- To assess different uncertainty scores
- To examine to what extent a SOTA model captures data uncertainty and model uncertainty
- To explore effects that influence uncertainty estimation in the perspectives of morphology, inventory organization and semantic relations

Reference

- [Conia and Navigli, 2021] Simone Conia and Roberto Navigli. 2021. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration. In EACL: Main Volume, pages 3269–3275.
- [Maru et al., 2022] Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. Nibbling at the hard core of word sense disambiguation.ACL (Volume 1: Long Papers), pages 4724–4737
- Alessandro Raganato, Jose Camacho-Collados, andRoberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In ACL: Volume 1, Long Papers, pages 99–110.
- Maru, Marco, et al. "Nibbling at the hard core of Word Sense Disambiguation." Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022.

Thank you for your attention!

For more information, please refer to:

https://github.com/RyanLiut/WSD-UE