Ambiguity Meets Uncertainty: Investigating Uncertainty Estimation for Word Sense Disambiguation

Zhu Liu
Tsinghua University
School of Humanities
liuzhu22@mails.tsinghua.edu.cn

Ying Liu
Tsinghua University
School of Humanities
yingliu@tsinghua.edu.cn

Abstract

Word sense disambiguation (WSD), which aims to determine an appropriate sense for a target word given its context, is crucial for natural language understanding. Existing supervised methods treat WSD as a classification task and have achieved remarkable performance. However, they ignore uncertainty estimation (UE) in the real-world setting, where the data is always noisy and out of distribution. This paper extensively studies UE on the benchmark designed for WSD. Specifically, we first compare four uncertainty scores for a state-of-the-art WSD model and verify that the conventional predictive probabilities obtained at the final layer of the model are inadequate to quantify uncertainty. Then, we examine the capability of capturing data and model uncertainties by the model with the selected UE score on well-designed test scenarios and discover that the model adequately reflects data uncertainty but underestimates model uncertainty. Furthermore, we explore numerous lexical properties that intrinsically affect data uncertainty and provide a detailed analysis of four critical aspects: the syntactic category, morphology, sense granularity, and semantic relations. The code is available at https://github.com/RyanLiut/WSD-UE.

1 Introduction

Disambiguating a word in a given context is fundamental to natural language understanding (NLU) tasks, such as machine translation (Gonzales et al., 2017), question answering (Ferrández et al., 2006), and coreference resolution (Hu and Liu, 2011). This task of word sense disambiguation (WSD) targets polysemous or homonymous words and determines the most appropriate sense based on their surrounding contexts. For example, the ambiguous word *book* refers to two completely distinct meanings in the following sentences: i)"*Book* a hotel, please.", ii) "Read the *book*, please". The phenomenon is universal to all languages and has been paid much attention since the very beginning of artificial intelligence (AI) (Weaver, 1952).

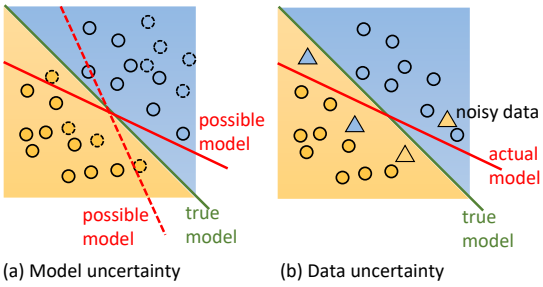

Figure 1: Two types of uncertainties in the case of classification. The green line indicates the true model (decision boundary), while the red shows possible models. Circles and triangles with different colors illustrate clean and noisy data with corresponding labels.

Existing supervised methods (Blevins and Zettlemoyer, 2020; Conia and Navigli, 2021; Bevilacqua and Navigli, 2020; Calabrese et al., 2021; Huang et al., 2019) cast WSD as a classification task in which a neural networks (NNs)-based classifier is trained from WordNet (Miller et al., 1990), a dictionary-like inventory. Although they have achieved the state of the art on WSD benchmarks, with some even breaking through the estimated upper bound on human inter-annotator agreement in terms of accuracy (Bevilacqua and Navigli, 2020), they do not capture or measure uncertainty. Uncertainty estimation (UE) answers a question as follows: *To what extent is the model certain that its choices are correct?* A model can be unsure due to the noisy or out-of-domain data, especially in a real-world setting. This estimation delivers valuable insights to the WSD practitioners since we could pass the input with high uncertainty to a human for classification.

UE is an essential requirement for WSD. Interestingly, the word "ambiguous" (in terms of the task of word sense *disambiguation*) itself is ambiguous: it refers to i) doubtful or uncertain especially from

Findings of the Association for Computational Linguistics: ACL 2023, pages 3963–3977
July 9-14, 2023 ©2023 Association for Computational Linguistics

obscurity or indistinctness, and ii) capable of being understood in two or more possible senses or ways, according to the Merriam-Webster dictionary[1]. The conventional treatment only considers its second aspect but disregards the first uncertainty-related sense. In reality, there are many situations where uncertainties arise (Yarin, 2016). The first situation assumes a true model to which each trained model approximates. Uncertainty appears when the structures and parameters of the possible models vary; we refer to it as model uncertainty (Figure 1 (a)) in this paper. *Model uncertainty* can be reduced when collecting enough data, i.e., adequate knowledge to recognize the true model and out-of-distribution (OOD) data is always used to test model uncertainty. It has been observed that WSD is prone to domain shift and bias towards the most frequent sense (MFS) (Raganato et al., 2017). Therefore, it is essential to quantify model uncertainty in the task.

Another uncertainty is related to the data itself and cannot be explained away, which is referred to as *data uncertainty* (also called aleatoric uncertainty). Data uncertainty happens when the observation is imperfect, noisy, or obscure (Figure 1 (b)). Even if there is enough data, we cannot obtain results with high confidence. WSD is context-sensitive, and the model output could be divergent due to partial or missing context. Even worse, some words have literal and non-literal meanings and can be understood differently. With a fine-grained WordNet (Miller et al., 1990) as a reference inventory, the inter-annotator disagreement is up to 20% to 30% (Navigli, 2009): even human annotators cannot agree on the correct sense of these words.

In this paper, we perform extensive experiments to assess the uncertainty of a SOTA model (Conia and Navigli, 2021) on WSD benchmarks. First, we compare the probability of the model output with the other three uncertainty scores and conclude that this probability is inadequate to UE, which is consistent with previous research (Gal and Ghahramani, 2016). Then, with the selected score, we evaluate data uncertainty in two designed scenarios: window-controlled and syntax-controlled contexts, which simulate noisy real-world data. Further, we estimate model uncertainty on an existing OOD dataset (Maru et al., 2022) and find that the model underestimates model uncertainty compared to the adequate measure of data uncertainty. Finally, we design an extensive controlled procedure to determine which lexical properties affect uncertainty estimation. The results demonstrate that morphology (parts of speech and number of morphemes), inventory organization (number of annotated ground-truth senses and polysemy degree) and semantic relations (hyponym) influence the uncertainty scores.

[1] https://www.merriam-webster.com/dictionary/ambiguous

2 Related Work

2.1 Word Sense Disambiguation

Methods of WSD are usually split into two categories, which are knowledge-based and supervised models. Knowledge-based methods employ graph algorithms, e.g., clique approximation (Moro et al., 2014), random walks (Agirre et al., 2014), or game theory (Tripodi and Navigli, 2019) on semantic networks, such as WordNet (Miller et al., 1990), BabelNet (Navigli and Ponzetto, 2012). These methods do not acquire much annotation effort but usually perform worse than their supervised counterpart due to their independence from the annotated data. Supervised disambiguation is data-driven and utilizes manually sense-annotated data sets. Regarding each candidate sense as a class, these models treat WSD as the task of multi-class classification and utilize deep learning techniques, e.g., transformers (Conia and Navigli, 2021; Bevilacqua and Navigli, 2019). Some also integrate various parts of the knowledge base, such as neighboring embeddings (Loureiro and Jorge, 2019), relations (Conia and Navigli, 2021), and graph structure (Bevilacqua and Navigli, 2020). These methods have achieved SOTA performance and even broken through the ceiling human could reach (Bevilacqua and Navigli, 2020). However, these methods treat disambiguation as a deterministic process and neglect the aspect of uncertainty.

2.2 Uncertainty Estimation

Uncertainty estimation (UE) has been studied extensively, especially in computer vision (Gal et al., 2017) and robust AI (Stutz, 2022). Methods capture uncertainty in a Bayesian or non-Bayesian manner. Bayesian neural networks (Neal, 2012) offer a mathematical grounded framework to model predictive uncertainty but usually comes with prohibitive inference cost. Recent work proved MC Dropout approximates Bayesian inference in deep Gaussian Processes and has been widely applied

in many UE applications (Vazhentsev et al., 2022; Kochkina and Liakata, 2020) due to its simplicity. During recent years, the field of natural language processing has witnessed the development of an increasing number of uncertain-aware applications, such as Machine Translation (Glushkova et al., 2021), Summarization (Gidiotis and Tsoumakas, 2021) and Information Retrieval (Penha and Hauff, 2021). Nevertheless, little attention has been paid to the combination of UE and WSD. An early work (Zhu et al., 2008) explored uncertainty to select informative data in their active learning framework. However, the uncertainty estimation for WSD is not explored extensively, as we do in a quantitative and qualitative way.

3 Uncertainty Scenarios

3.1 Problem Formulation

Given a target word w_i in a context $c_i = (w_0, w_1, ..., w_i, ..., w_W)$ of W words, a WSD model selects the best label $\hat{y}_i$ from a candidate sense set $S_i = (y_1, y_2, ..., y_M)$ consisting of M classes. A neural network p_θ with the parameter θ usually obtains a probability p_i over M classes by a softmax function which normalizes the model output f_i:

$$p_i = \text{SoftMax}(f_i(w_i|c_i;\theta)). \quad (1)$$

During training, the probability is used to calculate cross-entropy loss, which can be recognized as a probability for each candidate class during the inference. Such a point estimation of model function has been erroneously interpreted as model confidence (Gal and Ghahramani, 2016). The goal of UE is to find a suitable p_i to better reflect true predictive distribution under data and model uncertainty sources. Suppose we have a reasonable score $s(p_i) \in \mathcal{S}$ indicating UE, where $\mathcal{S}$ is a metric space, we expect $s^a > s^b$ when a situation a is more uncertain than b.

3.2 Data Uncertainty: Controllable Context

Data uncertainty measures the uncertainty caused by imperfect or noisy data. We consider that such noises could happen in the context surrounding the target word, considering WSD is a context-sensitive task. With different degrees of missing parts in the context, the model is expected to obtain predictions with different qualifications of uncertainty. To simulate this scenario, we control the range of context based on two signals: the window and the syntax, as illustrated in Figure 2.

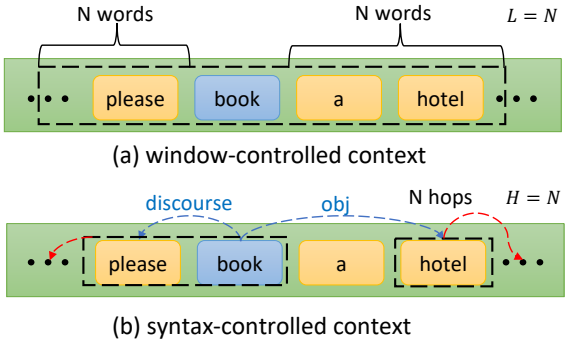

Figure 2: Two types of controlled context in the data uncertainty setting. The target word is highlighted in blue. The box with a black dotted line shows the final chosen context. We show the dependency relation in blue and red.

3.2.1 Window-controlled Context

We choose L words both on the left and right of the target word w_i as the window-controlled context $c_L^{\text{WC}} = (w_l, w_{i-1}, w_i, w_{i+1}, ..., w_h)$, where $l = \max(i - L, 0)$ and $h = \min(i + L, W)$ are the lower index and the higher index. With a hypothesis that longer context tends to contain more clues to disambiguate a word and a suitable UE score s, we expect that $s_a^{\text{WC}} > s_b^{\text{WC}}$, where two window-controlled contexts are extracted with the length of a and b, and $a < b$.

3.2.2 Syntax-controlled Context

In our second controlled method, we utilize the neighboring syntax around w_i. Specifically, we parse the universal syntactical dependency relations between words using tools of Stanza (Qi et al., 2020). This is represented as a form of graph structure $\mathcal{G} = (\mathcal{N}, \mathcal{R})$, where $\mathcal{N}$ denotes the nodes, i.e., each word, and $\mathcal{R} =< n^h, n^t, r >$ is the relation r from the head node n^h to tail node n^t. For example, when r is *nsubj*, that means n^h is the subject of n^t. We iteratively obtain a syntax-related [2] neighboring set with the H hops of the target word w_i as c_H^{DP} in the following approach. Initially, c_H^{DP} only contains w_i. After one hop, c_H^{DP} collects the head node and tail nodes of w_i. The procedure is repeated H times, with more syntactically related words added. We also rationally hypothesize a smaller s^{DP}, which measures uncertainty under

[2]We denote this scenario as DP, since we utilize dependency parsing as the syntactic representation.

syntax-controlled context, favors the context with a larger H. We highlight that the syntax-controlled context leverages the nonlinear dependency distance (Heringer et al., 1980) between words in connection, compared to the linear distance in the scenario of window-controlled context.

3.3 Model Uncertainty: OOD Test

Model uncertainty is another crucial aspect of UE, widely studied in the machine learning community. Lacking knowledge, models with different architectures and parameters could output indeterminate results. Testing a model on OOD datasets is a usual method to estimate model uncertainty. In the task of WSD, we employ an existing dataset 42D (Maru et al., 2022) designed for a more challenging benchmark. This dataset built on the British National Corpus is challenging because 1) for each instance, the ground truth does not occur in SemCor (Miller et al., 1994), which is the standard training data for WSD, and 2) is not the first sense in WordNet to avoid most frequent sense bias issue (Campolungo et al., 2022). 42D also has different text domains from the training corpus. These confirm that 42D is an ideal OOD dataset.

4 Experiments

4.1 Model and Datasets

We conduct our UE for a SOTA model MLS (Conia and Navigli, 2021), with the best parameters released by the authors. They framed WSD as a multi-label problem and trained a BERT-large-cased model (Kenton and Toutanova, 2019) on the standard WSD training dataset SemCor (Miller et al., 1994). We follow their settings except for using Dropout during inference when performing Monte Carlo Dropout (MC Dropout). We set the number of samples T to be 20, conduct 3 rounds, and report the averaged performance.

As regards the evaluation benchmark, we use the Unified Evaluation Framework for English all-words WSD proposed by (Raganato et al., 2017). This includes five standard datasets, namely, Senseval-2, Senseval-3, SemEval-2007, SemEval-2013, and SemEval-2015. The whole datasets concatenating all these data with different parts of speech (POS) are also evaluated. Note that in our second part, We use a portion of SemEval-2007 to investigate data uncertainty and 42D is used for model uncertainty.

4.2 Uncertainty Estimation Scores

We apply four methods as our uncertainty estimation (UE) scores. One trivial baseline (Geifman and El-Yaniv, 2017) regards the Softmax output p_i as the confidence values over classes $y = s \in S$. We calculate the uncertainty score based on the maximum probability as $u_{\text{MP}}(x) = 1 - \max_{s \in S} p(y = s|x)$.

The other three methods are based on MC Dropout, which has been proved theoretically as approximate Bayesian inference in deep Gaussian processes (Gal and Ghahramani, 2016). Specifically, we conduct T stochastic forward passes during inference with Dropout random masks and obtain T probabilities p_t. Following the work (Vazhentsev et al., 2022), we use the following measures:

- Sampled maximum probability (SMP) takes the sample mean as the final confidence before an MP is applied: $u_{\text{SMP}} = 1 - \max_{s \in S} \frac{1}{T} \sum_{t=1}^{T} p_t^s$, where p_t^s refers to the probability of belonging to class s at the $t'th$ forward pass.

- Probability variance (PV) (Gal et al., 2017) calculates the variance before averaging over all the class probabilities: $u_{\text{PV}} = \frac{1}{S} \sum_{s=1}^{S} \left(\frac{1}{T} \sum_{t=1}^{T} (p_t^s - \overline{p^s})^2 \right)$.

- Bayesian active learning by disagreement (BALD) (Houlsby et al., 2011) measures the mutual information between model parameters and predictive distribution: $u_{\text{BALD}} = -\sum_{s=1}^{S} \overline{p^s} \log \overline{p^s} + \frac{1}{T} \sum_{s,t} p_t^s \log p_t^s$.

Note that these scores are instance-specific and we report the averaged results over all the samples.

4.3 Metrics on UE scores

While UE scores are a measure of uncertainty, we also need metrics to judge and compare the quality of different UE scores. A hypothesis is that a sample with a high uncertainty score is more likely to be erroneous and removing such instances could boost the performance. We employ two metrics following the work (Vazhentsev et al., 2022): area under the risk courage curve (**RCC**) (El-Yaniv et al., 2010) and reversed pair proportion (**RPP**) (Xin et al., 2021). RCC calculates the cumulative sum of loss due to misclassification according to the uncertainty level for rejections of the predictions.

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.

A larger RCC indicates that uncertainty estimation negatively impacts the classification. Note that we use the normalized RCC by dividing the size of the dataset. RPP counts the proportion of instances whose uncertainty level is inconsistent with its loss level compared to another sample. For any pair of instances x_i and x_j with their UE score $u(x)$ and loss value $l(x)$:

$$RPP = \frac{1}{n^2} \sum_{i,j=1}^{n} \mathbb{1}[u(x_i) < u(x_j), l(x_i) > l(x_j)], \quad (2)$$

where n is the size of the dataset.

5 Results and Analysis

In the first part, we show the quantitative results of different UE scores and the performances of data and model uncertainty. Then a qualitative result demonstrates specific instances with a range of uncertainties. This motivates us to analyze which lexical properties mainly affect uncertainty in the last part.

5.1 Quantitative Results

5.1.1 Which UE score is better?

We measure the four UE scores, MP, SMP, PV, and BALD in terms of two metrics, RCC and RPP. The results of five standard datasets are shown in Table 1 while the performance on all the datasets involving different parts of speech is demonstrated in Table 2. For most of the data, SMP outperforms the other three scores in spite of some inconsistent results where MP has a slight advantage, such as on SemEval-15. Interestingly enough, softmax-based scores i.e., MP and SMP, surpass the other two, PV and BALD. Similar results can be observed in the work (Vazhentsev et al., 2022). This may be due to the fact that the former scores are directly used as the input of the maximum likelihood objective, thus more accurately approximating the real distribution.

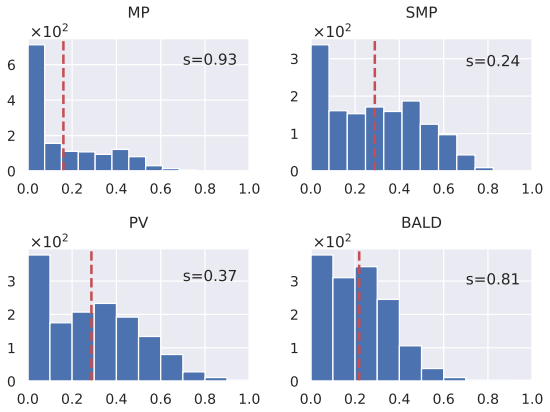

Figure 3: The distribution of four UE scores on misclassified instances of all datasets. A red dotted line indicates the average value. We calculate the sample skewness s for each score as well. Note that PV and BALD scores are normalized into the range from 0 to 1.

To further investigate the distribution of these four scores, we show the histograms of these scores in the misclassified instances, as illustrated in Figure 3. We also display the averaged value (a red dotted line) and the sample skewness s, calculated as the Fisher-Pearson coefficient (Zwillinger and Kokoska, 1999). Since here we focus on the misclassified samples, the cases of all the samples and those correctly classified are reported in Appendix A.1. This shows that MP has a more long-

tailed and skewed distribution than scores based on MC Dropout, indicating MP is overconfident towards the wrong cases. However, the other three metrics have a more balanced distribution. This verifies the common concern on the SoftMax output of a single forward as an indication of confidence.

Finally, given its outstanding performance, we chose SMP as our uncertainty score in the following experiments.

5.1.2 How does the model capture data uncertainty?

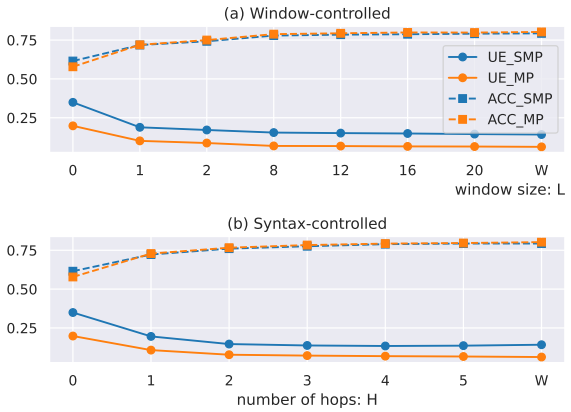

Figure 4: UE scores (SMP and MP) and accuracy (F1 score) vary depending on the range of context for (a) window-controlled setting and (b) syntax-controlled setting. Note that "0" indicates that only target words without context are available to the model. On the other hand, "W" means the whole context is available.

We verify data uncertainty in window-controlled and syntax-controlled scenarios, as shown in Figure 4. In the first setting, UE becomes less, and the accuracy grows with the increase of window size T. This indicates that the model perceives more and more confidence in the data, accessible to more neighboring words. The trend is similar in the syntax-controlled setting. These show that the model can adequately capture data uncertainty. SMP has a larger uncertainty than MP, especially in a sparse context, such as L or H is equal to 0 or 1, where the model is expected to be much more uncertain. We report the comparison of the other two sample-based scores, PV and BALD in Appendix A.2.

5.1.3 How does the model capture model uncertainty?

We examine the model uncertainty on the 42D dataset in Figure 5. The result shows OOD dataset

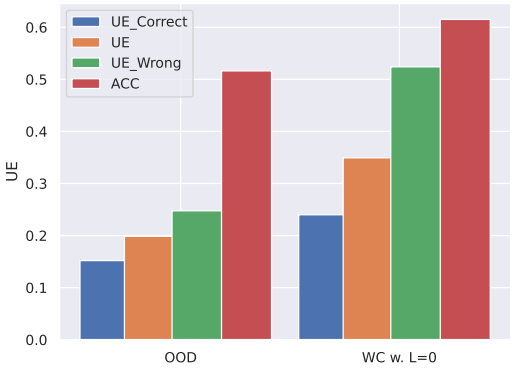

Figure 5: Uncertainty and accuracy (F1) scores for model uncertainty (OOD) and data uncertainty (controlled context) scenarios. We use window-controlled UE with L=0 (WC w. L=0). It is evaluated in all the data instances and wrongly (UE_Wrong) or correctly (UE_Correct) classified instances.

is indeed a challenging benchmark for WSD. However, even with worse performance, the model fails to give a high UE score. We compare it with the most uncertain cases but similar accuracy in the settings of data uncertainty, i.e., without any context when $L = 0$. The OOD setting has a lower level of uncertainty, especially in the misclassified samples, even if it has degraded performance. This implies that the model underestimates the uncertainty level in model uncertainty. We show the performance of MP, PV, and BALD in Appendix A.3.

5.2 Qualitative Results

To investigate what kinds of words given a context tend to be uncertain, we obtain the final UE score for each word by averaging SMP scores for instances sharing the same form of lemma. In Figure 6, We show the word clouds for words with the most uncertain (left (a)) and certain (right (b)) meanings. We remove some unrepresented words whose number of candidate senses is less than 3. With respect to the most uncertain lemmas, there are words such as *settle*, *cover* etc. Most of them are verbs and own multiple candidate senses. As for most certain cases, the senses of nouns like *bird*, *bed*, and *article* are determined with low uncertainty. These phenomena motivate us to investigate which lexical properties affect uncertainty estimation in the next part. It is noted that we concentrate on data uncertainty instead of model uncertainty, based on the investigation in Subsection 5.1, which appears due to the data itself, i.e., lexical character-

istics.

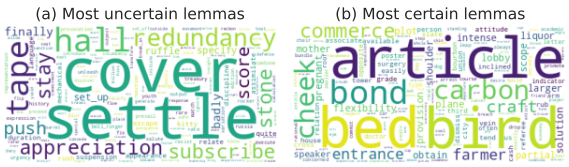

Figure 6: Word clouds for lemmas where a larger font indicates higher (a) or lower (b) UE scores.

5.3 Effects on Uncertainty

We explore which lexical properties affect uncertainty estimation from four aspects: the syntactic category (Folk and Morris, 2003), morphology [3](Lieber, 2004), sense granularity and semantic relations (Sternefeld and Zimmermann, 2013), motivated by linguistic and cognitive studies. Regarding syntactic categories, we focus on four i.e., parts of speech (**POS**) for target content words. Morphology aims at the number of morphemes (**nMorph**). A sense inventory refers to the sense items in a dictionary, whose granularity influences the candidate sense listing for the target word and its sense annotation (Kilgarriff, 1997). We consider two aspects:

- number of annotated ground-truth senses (**nGT**);
- number of candidate senses, i.e., polysemy degree (**nPD**);

To consider semantic interactions with other words, we utilize WordNet (Miller et al., 1990), a semantic network to extract lexical relations. Specifically, we concentrate on the hyponym and synonymy relations. A word (or sense) is a hyponym of another if the first is more specific, denoting a subclass of the other. For example, *table* is a hyponym of *furniture*. Each word as a node in WordNet lies in a hyponym tree, where the depth implies the degree of specification, denoted as **dHypo**. Meanwhile, we also explore the size of the synonymy set (**dSyno**) into which the ground-truth sense falls.

We perform linear regression analysis and conclude that most effects are significant as coefficients to the UE score, except for dSyno and ADV of POS.

[3]Here, we mainly consider derivational morphology. Multiword expressions e.g., compound words are included as well. Words with different inflectional morphology are regarded as the same lemma form.

This is consistent with our result in Subection 5.3.3. The summary of the linear regression is shown in Appendix A.4. Afterwards, we design a controlled procedure to analyze and balance different effects. First, samples are drawn from all the test instances depending on some conditions, including nGT and POS. Afterward, we aggregate test data in one of three manners: *instance* (I), *lemma* (L), and *sense* (S) and average the UE values for the instances with the same manner. I represents each occurrence of the target word, L considers words with different inflections (e.g., *works* and *worked*), and S targets words with the same ground-truth sense. The sampled data is then grouped into N levels in terms of the values for the different effects in question. Finally, we calculate the mean UE score for each group and their corresponding T-test and p values. We heuristically set different choices of N for different effects, considering the trade-off of level granularity and sample sparsity. The p-value is expected to be lower than 5%. The overall comparison is summarized in Table 3 with the number and value range of different levels in Table 4.

5.3.1 Syntactic Category and Morphology

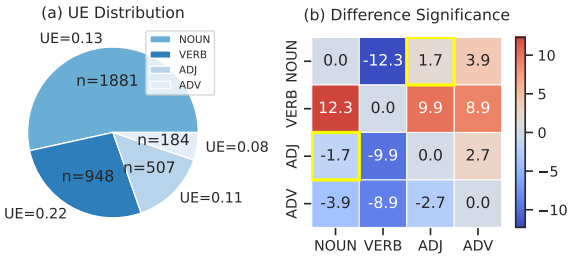

Figure 7: Averaged UE scores and numbers for instances aggregated by *sense*, with different parts of speech (a) and the corresponding difference significance for each pair (b). The heatmap (b) shows the T-test values where a higher absolute value (grids with a deeper color) indicates a more significant difference. We highlight the grid with a corresponding p value larger than 5%, implying no significant difference.

We show the averaged UE scores for instances with different POS and their corresponding T-test value in Figure 7. Except for the NOUN-ADJ pair, verbal instances are more significantly uncertain than NOUN or ADJ, while ADV has the least uncertainty. The result implies the senses of verbs are generally harder to determine than other categories, consistent with previous work (Barba et al., 2021; Campolungo et al., 2022). This is reflected in Table 2 and Figure 6.

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

Table 3: Different uncertainty estimations (SMP) for different levels and corresponding difference significance (p values) of various effects involving morphology, inventory organization and semantic relations. Agg. means aggregation manners of the lemma (L), instance (I), and sense (S).

Effect		L1	L2	L3
nMorph (N)	number	514	603	397
	range	(0,1.67]	(1.67,2]	(2,9]
nMorph (V)	number	200	313	132
	range	(0,2)	[2,2]	(2,6]
nMorph (A)	number	136	201	69
	range	(0,1.30]	(1.30,2]	(2,6]
nMorph (D)	number	25	85	36
	range	(0,2]	[2,2]	(2,6]
nGT	number	6913	340	-
	range	1	>1	-
nPD	number	1145	963	463
	range	(0,2]	(2,6]	(6,50]
dHypo	number	729	666	340
	range	(1,6]	(6,9]	(9,43]
dSyno	number	1109	1407	763
	range	(0,1]	(1,3]	(3,28]

Table 4: The number and range of effects quantified into different levels for various effects.

We further explore the effects of morphology in Table 3. After extracting morphemes for each word using an off-line tool [4], we count the number of morphemes (denoted as nMorph). Since words with different parts of speech may have distinct mechanisms of word formation rules, we split data according to POS before averaging their UE scores and calculating corresponding difference significance. It shows that generally, the more morphemes a word consists of, the more uncertain its semantics would be. This is expected from the perspective of derivational morphology since adding prefixes, or suffixes could specify the stem words and have a relatively predictable meaning. For example, "V-ation" indicates the action or process of the stem verb, e.g., education, memorization. According to T-test in Table 3, UE scores of different levels for nouns are significantly distinct, while the difference is not so significant for other categories. It is because the derivational nouns including compound words are more representative and productive than other categories. This can be demonstrated by the fact that nouns contain the highest number of morphemes as shown in Table 4.

[4] https://polyglot.readthedocs.io/en/latest

5.3.2 Sense Granularity

We first consider the number of ground-truth senses, i.e., nGT. During the annotation process, a not insignificant 5% of the target words is labeled multiple senses (Conia and Navigli, 2021). This reflects the difficulty in choosing the most appropriate meaning, even for human annotators. Given their contexts, the semantics of these words are expected to be more uncertain, and our result is consistent with this fact. We control nGT to be 1 in the remaining evaluation to eliminate its influence.

Second, we study the effect of polysemy degree (the number of possible candidates), i.e., nPD. It shows that target words with a more significant polysemy degree tend to be more uncertain. It is intuitively understandable because words with more possible meanings are always commonplace and easily prone to semantic change, e.g., *go*, *play*. Furthermore, their sense descriptions in WordNet are more fine-grained, indistinguishable in some cases even for humans. However, words with less polysemy degrees, such as compound words, are more certain in various contexts.

5.3.3 Semantic relation

We discuss the effects of semantic relations for the target word in terms of WordNet. We first consider the hyponym relations, i.e., the depth in which a word node lies in the hyponym relation tree, as denoted by dHypo. Since nouns have clearer instances of hyponymy relation, we only consider this category. The results displayed in Table 3 show that instances with a deeper hyponym tend to own a certain meaning and the difference between each pair of levels is significant. That indicates that more specific concepts have a more determinate disambiguation, which is intuitive.

Another semantic relation is synonymy, as represented by dSyno. The measurement reveals that instances among different levels of the number of synonyms do not differ from each other significantly. This implies that whether the ground-truth meaning has more neighbors with similar semantics has less impact on the decision of uncertainty.

6 Conclusion

We explore the uncertainty estimation for WSD. First, we compare various uncertainty scores. Then we choose SMP as the uncertainty indicator and examine to what extent a SOTA model captures data uncertainty and model uncertainty. Experiments demonstrate that the model estimates data uncertainty adequately but underestimates model uncertainty. We further explore effects that influence uncertainty estimation in the perspectives of morphology, inventory organization and semantic relations. We will integrate WSD with uncertainty estimation into downstream applications in the future.

7 Limitations

Despite being easily adapted to current deep learning architectures, one concern about multiple-forward sampling methods is efficiency, since it has to repeat T processes to evaluate uncertainty in the stage of inference. We leave efficient variants of sampling methods for future work.

Another glaring issue is the focus on only English. Different languages may have different effects on uncertainty estimation due to e.g., distinct forms of morphology. Thus, some conclusions may vary according to the language in question. We hope that follow-up works will refine and complement our insights on a more representative sample of natural languages.

8 Ethics Statement

We do not foresee any immediate negative ethical consequences of our research.

9 Broader Impact Statement

Knowing what we do not know, i.e., a well-calibrated uncertainty estimation, is fundamental for an AI-assisted application in the real world. In the area of word sense disambiguation, the ambiguity and vagueness inherent in lexical semantics require a model to represent and measure uncertainty effectively. Our work explores the combination of these two areas and hopes that it will provide an approach to understanding the characteristics of languages.

10 Acknowledgements

The authors thank the anonymous reviewers for their valuable comments and constructive feedback on the manuscript. We also thank Rui Fang for his discussions on the linear regression analysis. This work is supported by the 2018 National Major Program of Philosophy and Social Science Fund "Analyses and Researches of Classic Texts of Classical Literature Based on Big Data Technology" (18ZDA238) and Tsinghua University Initiative Scientific Research Program (2019THZWJC38).

References

Eneko Agirre, Oier López de Lacalle, and Aitor Soroa. 2014. Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, 40(1):57–84.

Edoardo Barba, Luigi Procopio, and Roberto Navigli. 2021. Consec: Word sense disambiguation as continuous sense comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1492–1503.

Michele Bevilacqua and Roberto Navigli. 2019. Quasi bidirectional encoder representations from transformers for word sense disambiguation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 122–131.

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017.

Agostina Calabrese, Michele Bevilacqua, and Roberto Navigli. 2021. Evilbert: Learning task-agnostic multimodal sense embeddings. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 481–487.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. Dibimt: A novel benchmark for measuring word sense disambiguation biases in machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4331–4352.

Simone Conia and Roberto Navigli. 2021. Framing word sense disambiguation as a multi-label problem for model-agnostic knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3269–3275.

Ran El-Yaniv et al. 2010. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(5).

S Ferrández, Sandra Roger, Antonio Ferrández, Antonia Aguilar, and Pilar López-Moreno. 2006. A new proposal of word sense disambiguation for nouns on a question answering system. *Advances in Natural Language Processing. Research in Computing Science*, 18:83–92.

Jocelyn R Folk and Robin K Morris. 2003. Effects of syntactic category assignment on lexical ambiguity resolution in reading: An eye movement analysis. *Memory & Cognition*, 31:87–99.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. *Advances in neural information processing systems*, 30.

Alexios Gidiotis and Grigorios Tsoumakas. 2021. Uncertainty-aware abstractive summarization. *arXiv preprint arXiv:2105.10155*.

Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André FT Martins. 2021. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3920–3938.

Annette Rios Gonzales, Laura Mascarell, and Rico Sennrich. 2017. Improving word sense disambiguation in neural machine translation with sense embeddings. In *Proceedings of the Second Conference on Machine Translation*, pages 11–19.

Hans Jürgen Heringer, Bruno Strecker, and Rainer Wimmer. 1980. *Syntax: Fragen, Lösungen, Alternativen*. Fink.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *stat*, 1050:24.

Shangfeng Hu and Chengfei Liu. 2011. Incorporating coreference resolution into word sense disambiguation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 265–276. Springer.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuan-Jing Huang. 2019. 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 3509–3514.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Elena Kochkina and Maria Liakata. 2020. Estimating predictive uncertainty for rumour verification models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6964–6981.

Rochelle Lieber. 2004. *Morphology and lexical semantics*, volume 104. Cambridge University Press.

Daniel Loureiro and Alipio Jorge. 2019. 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.

Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. Nibbling at the hard core of word sense disambiguation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4724–4737.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

George A Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G Thomas. 1994. Using a semantic concordance for sense identification. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial intelligence*, 193:217–250.

Radford M Neal. 2012. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Gustavo Penha and Claudia Hauff. 2021. On the calibration and uncertainty of neural learning to rank models for conversational search. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 160–170.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. 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.

Wolfgang Sternefeld and Thomas Ede Zimmermann. 2013. *Introduction to Semantics: An Essential Guide to the Composition of Meaning (Mouton Textbook)*. De Gruyter Mouton.

David Stutz. 2022. Understanding and improving robustness and uncertainty estimation in deep learning. *Saarländische Universitäts-und Landesbibliothek*.

Rocco Tripodi and Roberto Navigli. 2019. Game theory meets embeddings: a unified framework for word sense disambiguation. 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 88–99.

Artem Vazhentsev, Gleb Kuzmin, Artem Shelmanov, Akim Tsvigun, Evgenii Tsymbalov, Kirill Fedyanin, Maxim Panov, Alexander Panchenko, Gleb Gusev, Mikhail Burtsev, et al. 2022. Uncertainty estimation of transformer predictions for misclassification detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8237–8252.

Warren Weaver. 1952. Translation. In *Proceedings of the Conference on Mechanical Translation*.

Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. 2021. The art of abstention: Selective prediction and error regularization for natural language processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051.

Gal Yarin. 2016. Uncertainty in deep learning. *University of Cambridge, Cambridge*.

Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. 2008. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 1137–1144.

Daniel Zwillinger and Stephen Kokoska. 1999. *CRC standard probability and statistics tables and formulae*. Crc Press.

A Appendix

A.1 Distribution of UE scores

We illustrate the distribution of UE scores, i.e., MP, SMP, PV and BALD for all the test samples in Figure 8 and samples that are correctly predicted in Figure 9. We assume samples that the model could accurately predict are easy and thus have a more certain meaning. Although SMP is not so long-tailed as MP in the case of correctly predicted samples, we do not expect a metric "overconfident" in all the cases, especially in the misclassified instances.

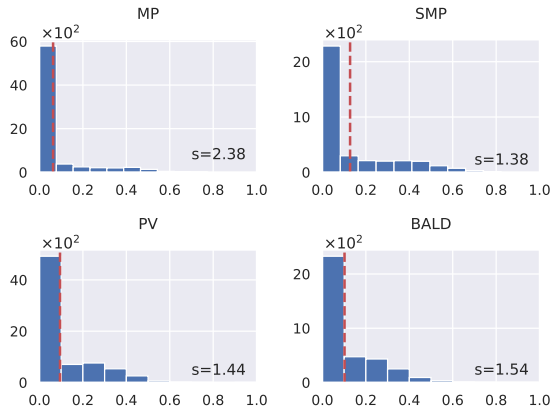

Figure 8: The distribution of four UE scores on all the test samples. The averaged value is indicated by a red dotted line. We calculate the sample skewness for each score as well.

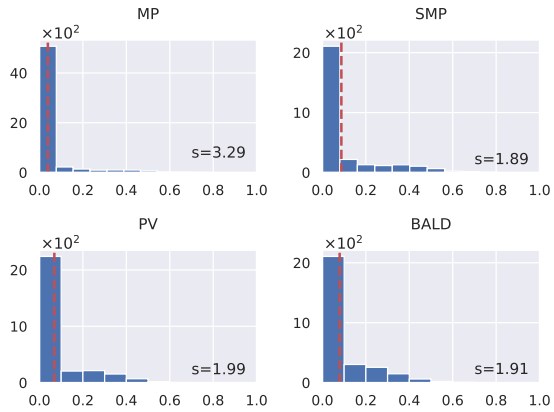

Figure 9: UE distribution on well-classified samples.

A.2 Other Scores for Data Uncertainty

We display the other two sample-based scores PV and BALD, in comparison with SMP in two data uncertainty scenarios in Figure 10. SMP has a higher uncertain score than the other two, especially in the more sparse context (e.g., $L = 0$), as we expected.

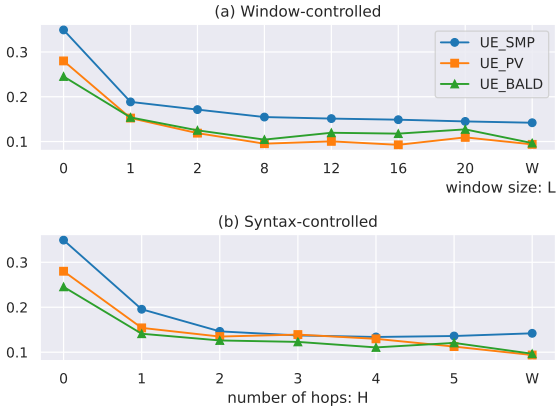

Figure 10: UE scores (SMP, PV, and BALD) vary depending on the range of context for (a) window-controlled setting and (b) syntax-controlled setting.

A.3 Other Scores for Model Uncertainty

We illustrate the other three UE scores (MP, PV and BALD) and accuracy for the scenario of model uncertainty compared with the least uncertain case for data uncertainty (L=0) in Figure 11, Figure 12 and Figure 13, respectively. The conclusion that UE scores underestimate model uncertainty is similar to that of MP.

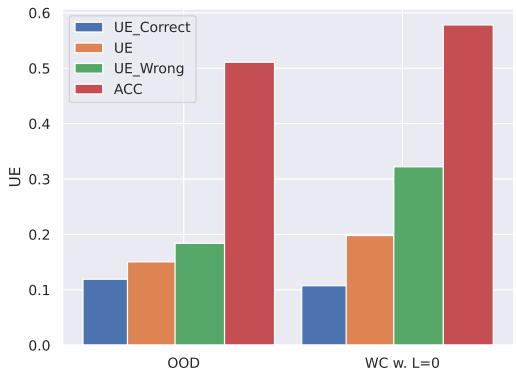

Figure 11: Uncertainty (MP) and accuracy scores for model uncertainty (OOD) and data uncertainty (controlled context) scenarios. We use window-controlled UE with L=0 (WC w. L=0). It is evaluated in all the data instances and wrongly (UE_Wrong) or correctly (UE_Correct) classified instances.

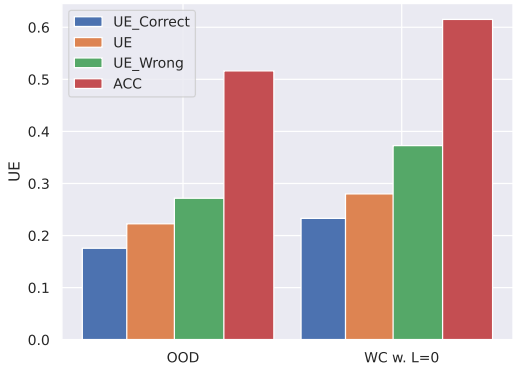

Figure 12: Uncertainty (PV) and accuracy scores for model uncertainty (OOD) and data uncertainty (controlled context) scenarios.

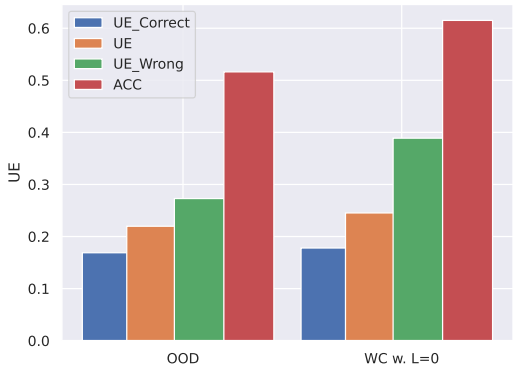

Figure 13: Uncertainty (BALD) and accuracy scores for model uncertainty (OOD) and data uncertainty (controlled context) scenarios.

A.4 Linear Regression Analysis

Figure 14 reports all the effects and corresponding coefficients and p-values of the linear regression model described in Subsection 5.3.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.58641 -0.10545 -0.06753  0.09504  0.53066

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0083035  0.0170900   0.486  0.62709
POSADV      -0.0175029  0.0142610  -1.227  0.21978
POSNOUN      0.0332515  0.0116023   2.866  0.00418 **
POSVERB      0.0687057  0.0098485   6.976 3.61e-12 ***
nMorph      -0.0115582  0.0035480  -3.258  0.00113 **
nGT          0.0843417  0.0120718   6.987 3.35e-12 ***
nPD          0.0086235  0.0004789  18.006  < 2e-16 ***
dHypo       -0.0021911  0.0011069  -1.979  0.04785 *
dSyno       -0.0012973  0.0014049  -0.923  0.35585
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1648 on 3511 degrees of freedom
Multiple R-squared:  0.175,   Adjusted R-squared:  0.1731
F-statistic:  93.1 on 8 and 3511 DF,  p-value: < 2.2e-16
```

Figure 14: Linear regression model predicting the UE score (SMP) by various effects.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☒ A2. Did you discuss any potential risks of your work?
Our paper does not see any risks.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

3,4

☑ B1. Did you cite the creators of artifacts you used?
3,4

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
3,4

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
4

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The benchmark is open and has been checked by many other reasearchers.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
4

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
4,5

C ☑ Did you run computational experiments?

4,5

☒ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
The model is proprosed by other work where the authors have reported that.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
3

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.