清華大學
Tsinghua University

Paper Sharing

Zhu Liu

2022.10.28

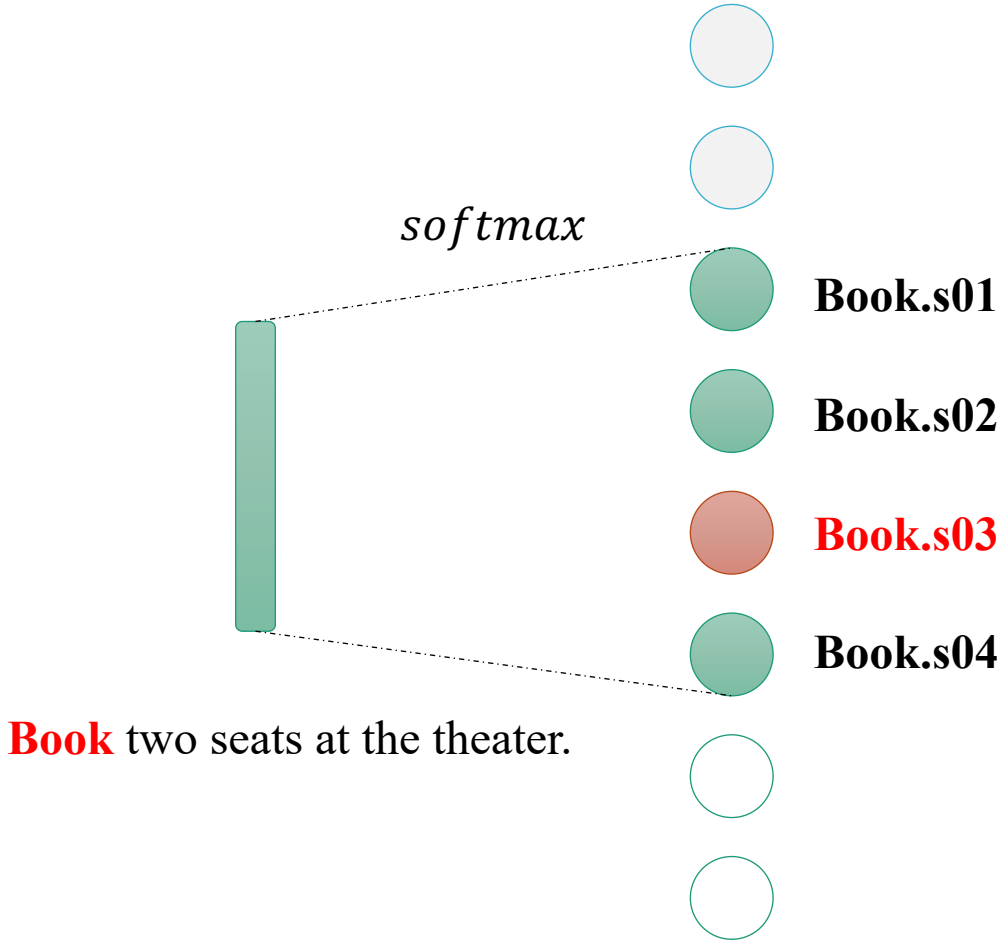
Framing Word Sense Disambiguation as a Multi-Label Problem for Model-Agnostic Knowledge Integration

Simone Conia **Roberto Navigli**
Sapienza NLP Group
Department of Computer Science
Sapienza University of Rome
{conia,navigli}@di.uniroma1.it

EACL 2021; Cited by 15

Motivation

- Most SOTA models treat WSD as a single-label classification task.
- E.g., BEM, EWISER, SVC, GlossBERT
- The likelihood distribution (the softmax output) is one multinomial distribution (多项分布) .



$$\mathcal{L}_{\mathrm{CE}}(w, \hat{s}) = -\log(y_{\hat{s}})$$

Motivation

- Most SOTA models treat WSD as a single-label classification task.
- Due to the fine-grained sense labelling (e.g. for Wordnet), the sense boundary for humans is not clear:

  1) ~20% disagreement

  2) **~5% multiple labels**

#senses	1	2	3
#instances	6913	322	18

4.69% multiple labels for Evaluation Dataset

Sense granularity

Verb

- S: (v) **book** (engage for a performance) *"Her agent had booked her for several concerts in Tokyo"*
- S: (v) reserve, hold, **book** (arrange for and reserve (something for someone else) in advance) *"reserve me a seat on a flight"; "The agent booked tickets to the show for the whole family"; "please hold a table at Maxim's"*
- S: (v) **book** (record a charge in a police register) *"The policeman booked her when she tried to solicit a man"*
- S: (v) **book** (register in a hotel booker)

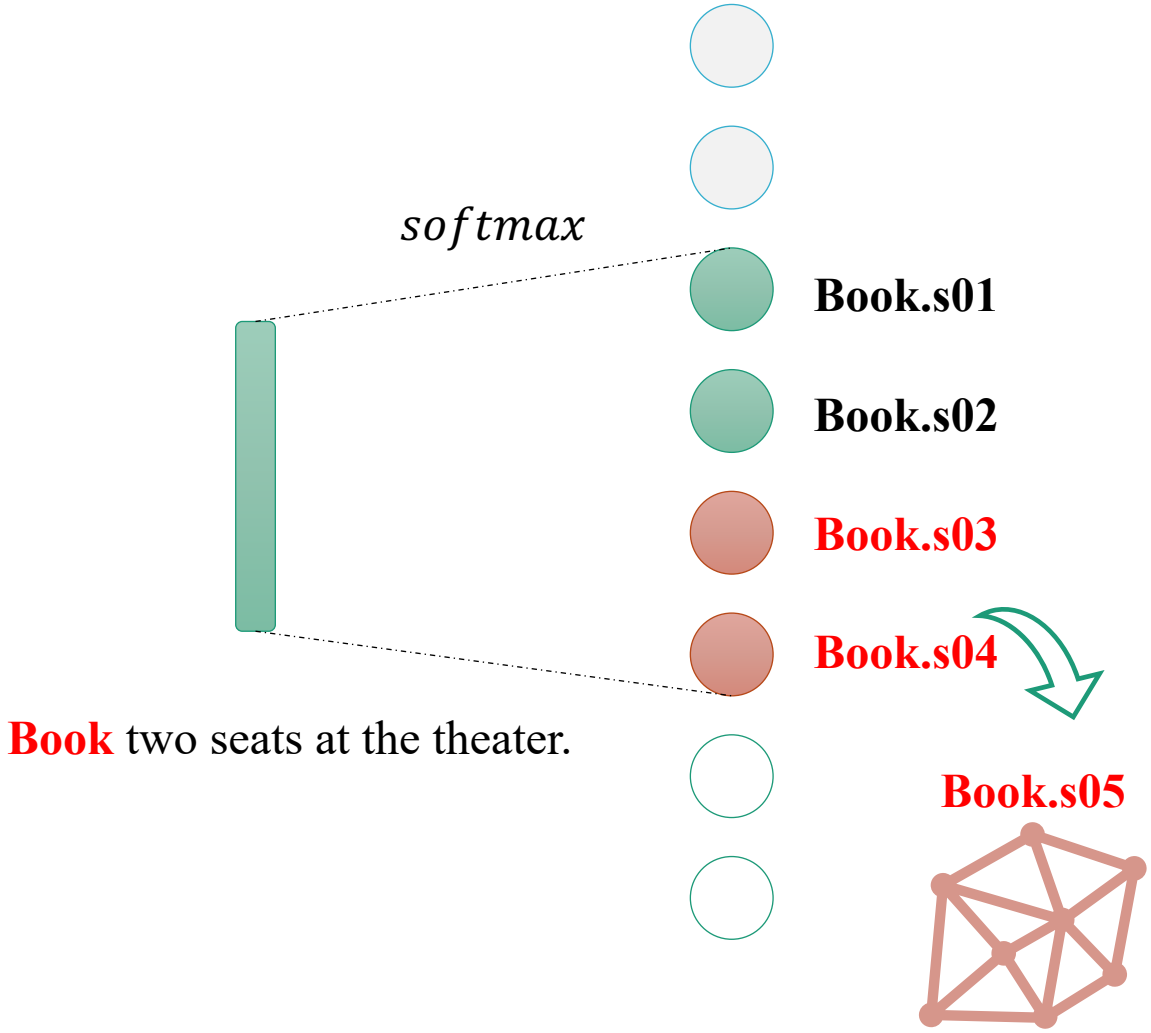
Search wordnet for “book”

Other views

- *The division of a word's meaning into senses is forced onto lexicographers by the economic and cultural setting within which they work. Lexicographers are obliged to describe words as if all words had a **discrete, non-overlapping** set of senses. It does not follow that they do, nor that lexicographers believe that they do. [Kilgrarriff, 1997, I don't believe in word senses]*
- *…produce **graded** ratings instead of making a binary choice. [Katrin Erk, 2009]*
- …… are not to be thought of as like five kinds of cheese, among which one must choose, but more like a casserole (砂锅) which has some pieces of clearly distinct identifiable content, but **a lot of stuff of uncertain and mixed origin in between**. ([Foundations of Statistical Natural Language Processing])

Contribution

- Re-frame the task as multi-label classification
- Data augmentation: Integrate senses from relational knowledge



Method

- From single-label to multi-label
- Cross entropy (CE) on one multinominal distribution → CE on multiple binomial distribution (BCE)

$$\mathcal{L}_{\mathrm{CE}}(w, \hat{s}) = -\log(y_{\hat{s}})$$

$$\mathcal{L}_{\mathrm{BCE}}(w, \hat{S}_w) = -\sum_{\hat{s} \in \hat{S}_w} \log(y_{\hat{s}}) - \sum_{s \in S_w \setminus \hat{S}_w} \log(1 - y_s)$$

$\hat{S}_w \subseteq S_w$ is the set of appropriate senses for the target word w in the given context c.

Method

- Data augmentation using structured knowledge
- G = <S, R>: S: sense nodes + R: semantic relations
- R: similar-to, derivationally-related, hypernymy or hyponymy …
- Enlarge the GT label set S_w to S_w^+ by adding one-hop neighboring sense.

$$\mathcal{L}_{\mathrm{BCE}}(w, \hat{S}_w^+) = -\sum_{\hat{s} \in \hat{S}_w^+} \log(y_{\hat{s}}) - \sum_{s \in S_w^+ \setminus \hat{S}_w^+} \log(1 - y_s) \qquad (3)$$

$$\text{where } \hat{S}_w^+ = \hat{S}_w \cup \{s_j : (\hat{s}_i, s_j) \in R, \hat{s}_i \in \hat{S}_w\}.$$

Data-based or knowledge-based

- A new method to incorporate knowledges
- [EWISE, 19] (sense compression);
  [EWISER, 20] (neighboring meanings);
  [SVC, 19] (gloss)
- Additional bonus: good generalization

- Experimental tip: Adding more related data (even noisy)

Experiments & Results

							Concatenation of ALL datasets			
		SE2	SE3	SE07	SE13	SE15	Nouns	Verbs	Adj	Adv	ALL
SemCor only	Raganato et al. (2017a)	72.0	69.1	64.8	66.9	71.5	71.5	57.5	75.0	83.8	69.9
	$\text{BERT}_{\text{Large}}$	76.3	73.2	66.2	71.7	74.1	–	–	–	–	73.5
	Hadiwinoto et al. (2019)	75.5	73.6	68.1	71.1	76.2	–	–	–	–	73.7
	Peters et al. (2019)	–	–	–	–	–	–	–	–	–	75.1
	Vial et al. (2019)	–	–	–	–	–	–	–	–	–	75.6
	Vial et al. (2019) - *Ensemble*	77.5	77.4	69.5	76.0	**78.3**	79.6	65.9	79.5	85.5	76.7
	This work	**78.4**	**77.8**	**72.2**	**76.7**	78.2	**80.1**	**67.0**	**80.5**	**86.2**	**77.6**
SemCor + definitions / examples	Loureiro and Jorge (2019)	76.3	75.6	68.1	75.1	77.0	78.0	64.0	80.7	84.5	75.4
	Scarlini et al. (2020a)	–	–	–	78.7	–	80.4	–	–	–	–
	Conia and Navigli (2020)	77.1	76.4	70.3	76.2	77.2	78.7	65.6	81.1	84.7	76.4
	Bevilacqua et al. (2020)	78.0	75.4	71.9	77.0	77.6	79.9	64.8	79.2	86.4	76.7
	Huang et al. (2019)	77.7	75.2	72.5	76.1	80.4	–	–	–	–	77.0
	Scarlini et al. (2020b)	78.0	77.1	71.0	77.3	83.2	80.6	68.3	80.5	83.5	77.9
	Blevins and Zettlemoyer (2020)	79.4	77.4	74.5	79.7	81.7	81.4	68.5	83.0	**87.9**	79.0
	Bevilacqua and Navigli (2020)	**80.8**	**79.0**	75.2	80.7	81.8	**82.9**	69.4	82.9	87.6	80.1
	This work	80.4	77.8	**76.2**	**81.8**	**83.3**	**82.9**	**70.3**	**83.4**	85.5	**80.2**

Table 1: WSD results in F_1 scores on Senseval-2 (SE2), Senseval-3 (SE3), SemEval-2007 (SE07), SemEval-2013 (SE13), SemEval-2015 (SE15), and the concatenation of all the datasets (ALL). Top: closed setting (only SemCor allowed as the training corpus without definitions and/or examples). Bottom: open setting (WordNet glosses and examples are also used for training).

Experiments & Results

WSD	Sim	See	Rel	Vrb	Hpe	Hpo	Hpe_I	Hpo_I	**SE07**	**ALL**
SL	–	–	–	–	–	–	–	–	69.0	74.7
ML	–	–	–	–	–	–	–	–	69.2	75.7
ML	✔	✔	✔	✔	–	–	–	–	70.6	76.6
ML	✔	✔	✔	✔	✔	–	–	–	71.0	77.0
ML	✔	✔	✔	✔	–	✔	–	–	**72.5**	77.4
ML	✔	✔	✔	✔	✔	✔	–	–	72.2	**77.6**
ML	✔	✔	✔	✔	✔	✔	✔	✔	72.2	**77.6**

Table 2: WSD results in F_1 scores on SemEval-2007 (SE07) and the concatenation of all the datasets (ALL). SL/ML: single-label/multi-label. Sim: similar-to. See: also-see. Rel: derivationally-related-forms. Vrb: verb-groups. Hpe: hypernymy. Hpo: hyponymy. Hpe_I: instance-hypernyms. Hpo_I: instance-hyponyms.

Conclusion

- Multi-label classification considers the granularity issue of sense inventories, and is simple and model-agnostic.
- Richer knowledge can be integrated in the new framework, with additional bonus, such as alleviating (knowledge acquisition bottle) KAB issue and generalized risks.

- [PERSONAL] The fine-grained inventory implies a soft, graded, and continuous sense representation. It is also a source of disagreed labeling (data uncertainty).

Reference

- Kilgarriff, Adam. "I don’t believe in word senses." *Computers and the Humanities* 31.2 (1997): 91-113.
- Erk, Katrin, and Diana McCarthy. "Graded word sense assignment." *Proceedings of the 2009 conference on empirical methods in natural language processing*. 2009.
- Bevilacqua, Michele, and Roberto Navigli. "Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.
- Kumar, Sawan, et al. "Zero-shot word sense disambiguation using sense definition embeddings." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.
- Vial, Loïc, Benjamin Lecouteux, and Didier Schwab. "Sense vocabulary compression through the semantic knowledge of wordnet for neural word sense disambiguation." *arXiv preprint arXiv:1905.05677* (2019).

Uncertainty Estimation of Transformer Predictions for Misclassification Detection

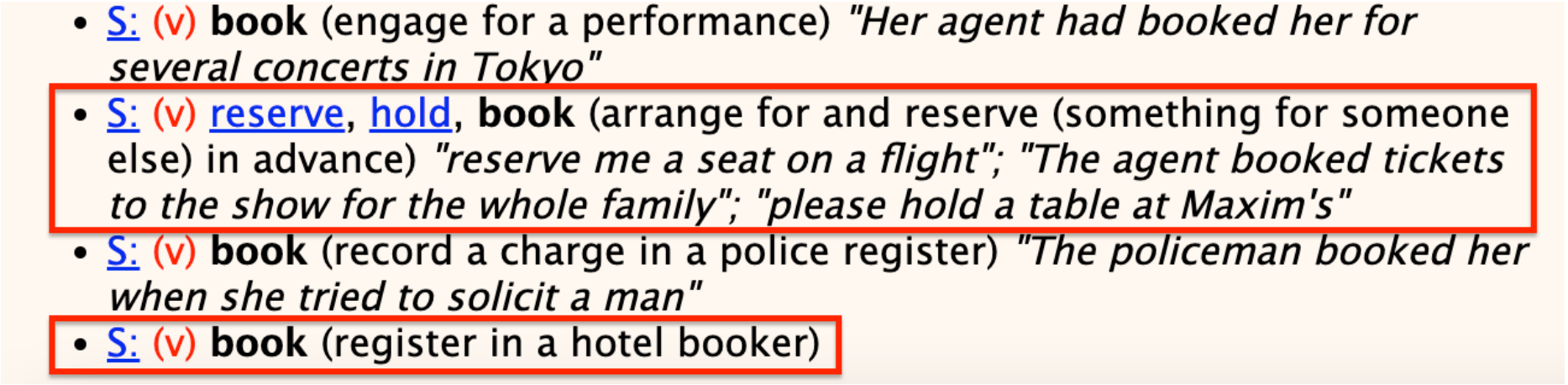
Artem Vazhentsev[1,2] ♢, Gleb Kuzmin[1,6] ♢, Artem Shelmanov[1,7] ♢, Akim Tsvigun[1,4], Evgenii Tsymbalov[2], Kirill Fedyanin[2], Maxim Panov[2], Alexander Panchenko[2], Gleb Gusev[1,3,5], Mikhail Burtsev[1,3], Manvel Avetisian[1,5], and Leonid Zhukov[1,4]

[1]AIRI, [2]Skoltech, [3]MIPT, [4]HSE, [5]Sber AI Lab, [6]FRC CSC RAS,
[7]ISP RAS Research Center for Trusted Artificial Intelligence

ACL‘2022

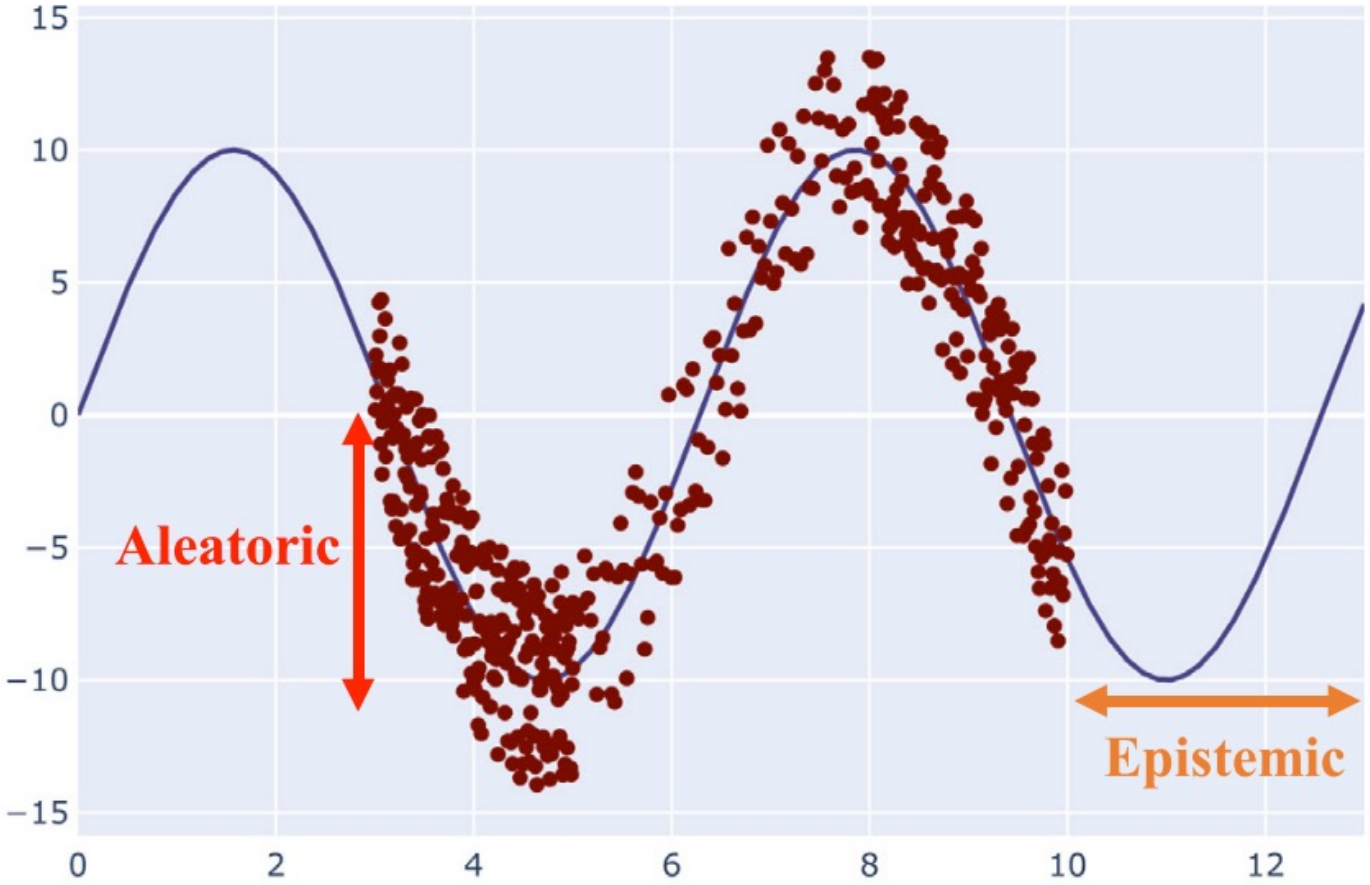
Preliminary – Uncertainty Estimation

- Models can not provide a deterministic answer due to various sources.
- A decision with a reliable uncertainty estimation (UE) is preferable in many fields, like medical diagnoses, autonomous driving, etc.
- Aleatoric (data) uncertainty: imperfect (noisy) data; irreducible
- Epistemic (model) uncertainty: OOD test data; reducible



0.8

Preliminary – Uncertainty Estimation



Preliminary – methods on UE

(There must be something stochastic/diverse.)

- **Bayesian models**

  The model weights are stochastic.

$$p(w|y,X) = \frac{p(y|X,w)p(w)}{p(y|X)}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

- **Non-Bayesian models**

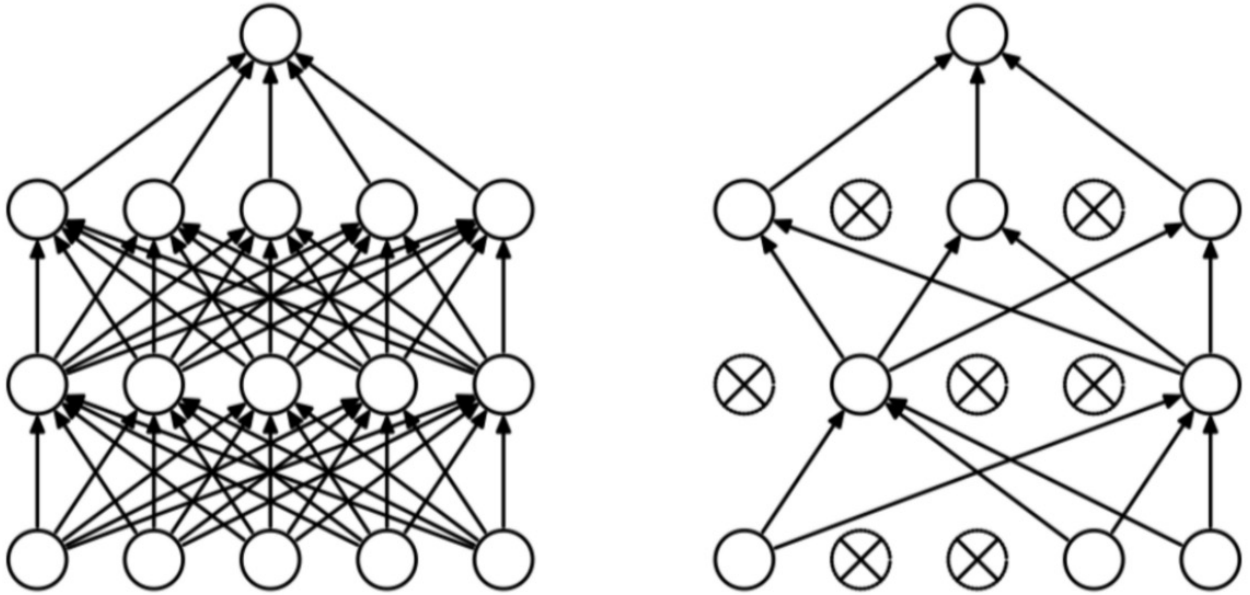
  Softmax response: unreliable “confidence”

  Gaussian process: A gaussian likelihood and prior

  Ensemble models: different models

  Mento Carlo Dropout: stochastically activated hidden states.

Preliminary - (MC) Dropout



(a) Standard Neural Net (b) After applying dropout.

Dropout: randomly “discard” (mask) some nodes in terms of a Bernoulli distribution during training but freeze during inference. (A useful technique to avoid **over-fitting**)
MC Dropout: Stochastic forward passes during reference. [Gal and Ghahramani, 2016]

Introduction

- Uncertainty estimation (UE) is crucial for active learning, adversarial attack detection, OOD instance detection, and **misclassification detection**.
- To investigate UE on transformer architecture in two classification tasks: text classification and named entity recognition (NER).
- A hidden hypothesis: an instance with high uncertainty is harmful, hard to decide. We need to detect them and pass them to human for classification.

Contributions

- Two novel computationally cheap methods

(1) For the stochastic model, the paper modifies the way of sampling a dropout mask in MC dropout.

(2) For the deterministic model, the method leverages Mahalanobis distance as a metric to UE with spectral normalization.

- The first to investigate UE methods on the NER task.
- An extensive empirical evaluation combined with some regularization techniques.

Background and Methods

How to represent/estimate uncertainty? (diversity, variations, change, entropy…)

- Softmax Response: $u_{\mathrm{SR}}(x) = 1 - \max_{c \in C} p(y = c|x).$

(We can also use entropy: $-\mathrm{E}[p \log p]$)

NOTE: SR is just a trivial baseline for UE due to its unreliability.

Background and Methods

- Standard Monta Carlo Dropout (MC dropout)

1) Sampled maximum probability (SMP)

$$u_{\mathrm{SMP}} = 1 - \max_{c \in C} \frac{1}{T} \sum_{t=1}^{T} p_t^c,$$

2) Probability variance (PV)

$$u_{\mathrm{PV}} = \frac{1}{C} \sum_{c=1}^{C} \left(\frac{1}{T} \sum_{t=1}^{T} (p_t^c - \overline{p^c})^2 \right) \qquad \overline{p^c} = \frac{1}{T} \sum_t p_t^c$$

3) Bayesian active learning by disagreement (BALD)

$$u_{\mathrm{BALD}} = -\sum_{c=1}^{C} \overline{p^c} \log \overline{p^c} + \frac{1}{T} \sum_{c,t} p_t^c \log p_t^c$$

NOTE that standard MC dropout needs activate all the dropout layers.

Background and Methods

- **Determinantal Point Process** Monte Carlo Dropout (DPP MC dropout)
- DPP could enhance the diversity of sampled set.
- Another work [Shelmanov, Artem, et al.] combined MC dropout with DPP for sampling neurons in a dropout layer. It shows that using stochasticity in the last dropout layer only is enough.

$$M_h^{DPP} \sim DPP(C_h) \quad P\left[M_h^{DPP} = S\right] = \frac{\det(C_h^S)}{\det(C_h + I)}$$

C_h is the similarity matrix between neurons; C_h^S is the square submatrix of C_h.

Background and Methods (Ours)

- After sampling a DPP mask pool, the paper proposed two strategies to continue to sample from the pool.

1) DDPP (+DPP)

   By applying DPP sampling again to the pool of pre-generated masks.

2) DDPP (+OOD)

   By selecting the masks that yield the highest PV scores on the given OOD dataset

Methods (Deterministic UE)

- Another work [Lee, Kimin, et al.] suggests measuring UE by the Mahalanobis distance between a test instance and the closest class-conditional Gaussian distribution:

$$u_{\text{MD}} = \min_{c \in C} (h_i - \mu_c)^T \Sigma^{-1} (h_i - \mu_c),$$

where h_i is the hidden representation of instance i, μ and Σ are the statistics for all the training instances.

- (Ours) To utilize the distance-preserving representation by **spectral normalization** on the weights of classification head.

$$\tilde{W} = \frac{W}{\nu}$$

Training Loss Regularizations

- Confidence Error Regularizer (CER)

  Penalty for instances with larger error but bigger confidence [Xin et al.]

$$L_{reg} = \sum_{i,j=1}^{k} \Delta_{i,j} \mathbb{1}[e_i > e_j],$$

$$\Delta_{i,j} = \max\{0, \max_c p_i^c - \max_c p_j^c\}^2,$$

- Metric Regularizer

  Shorten the intra-class distance and enlarge the inter-class distance [Zhang, Xuchao, et al.]

$$L_{reg} = \sum_{c=1}^{C} \left\{ L_{intra}(c) + \varepsilon \sum_{k \neq c} L_{inter}(c,k) \right\}, \quad (10)$$

$$L_{intra}(c) = \frac{2}{|S_c|^2 - |S_c|} \sum_{i,j \in S_c, i<j} D(h_i, h_j), \quad (11)$$

$$L_{inter}(c,k) = \frac{1}{|S_c| \cdot |S_k|} \sum_{i \in S_c, j \in S_k} [\gamma - D(h_i,h_j)]_+, \quad (12)$$

$$D(r_i, r_j) = \frac{1}{d} ||h_i - h_j||_2^2, \quad (13)$$

Experiment setup

- Dataset

1) Text classification: three datasets from GLUE benchmark

   MRPC (Microsoft Research Paraphrase Corpus)

   CoLA (Corpus of Linguistic Acceptability)

   SST-2 (Stanford Sentiment Treebank)

2) NER: CoNLL-2003 task

Metrics (on UE)

- A core hypothesis: samples with high uncertainty (low confidence) should have a low loss (error).

(1) RCC-AUC (area under the risk coverage curve)

Sum of loss due to misclassification (cumulative risk) depending on the uncertainty level used for rejection of predictions.

(2) RPP (reversed pair proportion)

instances with higher confidence

should have a lower loss.

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

(3) Accuracy rejection curve (ARC)

All rejected instances by certain uncertainty level are labeled with an oracle

UE v. metrics on UE

- UE metrics: SMP, PV, BALD, MD
- How to measure UE:

1) Indirect - Accuracy based on UE metrics (more about epistemic uncertainty), including RCC-AUC, RPP, ARC

2) direct: is the data *itself* uncertain? (aleatoric uncertainty)

The paper lacks this measure. Maybe it can be measured by human correlation!

Results

Method	Reg. Type	UE Score	MRPC		SST-2		CoLA		CoNLL-2003 (token level)		CoNLL-2003 (seq. level)	
			RCC-AUC ↓	RPP ↓	RCC-AUC ↓	RPP ↓	RCC-AUC ↓	RPP ↓	RCC-AUC ↓	RPP ↓	RCC-AUC ↓	RPP ↓
MC	-	PV	13.97±1.16	1.68±0.09	12.90±1.92	0.82±0.11	44.35±4.90	2.06±0.16	6.32±1.66	0.10±0.02	16.05±3.78	1.93±0.43
MC	-	BALD	14.21±1.04	1.69±0.09	12.98±1.87	0.82±0.10	45.06±4.90	2.08±0.17	6.44±1.86	0.10±0.02	16.28±4.00	1.96±0.45
MC	-	SMP	14.38±2.07	1.76±0.19	14.00±2.20	0.91±0.15	42.95±5.98	2.01±0.15	**6.04±1.03**	**0.09±0.02**	**15.79±3.34**	1.80±0.35
MC	CER	PV	**12.82±1.89**	**1.60±0.13**	12.18±1.20	0.80±0.10	46.84±9.19	2.11±0.23	6.92±1.22	0.10±0.02	17.05±3.14	1.91±0.36
MC	CER	BALD	12.89±1.89	**1.60±0.13**	12.39±1.23	0.81±0.09	47.34±8.30	2.14±0.24	7.16±1.15	0.11±0.02	17.25±3.05	1.93±0.35
MC	CER	SMP	12.91±2.15	1.67±0.15	12.22±1.31	0.82±0.09	46.10±11.07	2.05±0.22	6.69±1.38	0.10±0.02	16.81±1.61	1.81±0.14
MC	metric	PV	14.21±1.95	1.73±0.23	12.28±1.77	0.80±0.11	42.35±0.69	2.04±0.07	6.69±0.89	0.10±0.01	17.17±1.90	1.93±0.31
MC	metric	BALD	14.55±2.31	1.73±0.23	**12.08±1.79**	**0.79±0.10**	43.76±0.55	2.08±0.07	6.91±1.02	0.10±0.01	17.47±1.85	1.98±0.30
MC	metric	SMP	13.39±1.19	1.72±0.20	13.55±1.65	0.90±0.14	**40.88±1.25**	**2.01±0.09**	6.30±0.98	0.10±0.01	16.81±1.40	**1.80±0.23**
DDPP (+DPP) (ours)	-	PV	22.30±7.15	2.58±0.65	16.70±1.38	1.12±0.12	49.75±3.96	2.44±0.29	6.12±0.71	0.10±0.01	16.78±2.44	1.93±0.20
DDPP (+DPP) (ours)	-	BALD	23.08±7.00	2.63±0.63	16.08±2.37	1.05±0.18	49.59±5.40	2.48±0.31	6.39±0.64	0.10±0.01	21.53±4.77	2.63±0.45
DDPP (+DPP) (ours)	-	SMP	21.79±7.72	2.57±0.68	17.55±3.03	1.19±0.23	47.86±5.51	2.39±0.31	6.08±0.62	0.10±0.01	17.71±2.77	2.05±0.23
DDPP (+DPP) (ours)	CER	PV	15.12±2.27	2.03±0.24	13.56±1.37	0.91±0.14	54.51±8.80	2.58±0.22	6.98±0.98	0.11±0.02	19.44±1.15	2.13±0.17
DDPP (+DPP) (ours)	CER	BALD	15.94±3.77	2.07±0.36	14.87±2.22	0.96±0.13	55.11±7.42	2.61±0.31	7.90±1.95	0.12±0.01	26.20±6.41	3.11±0.56
DDPP (+DPP) (ours)	CER	SMP	14.75±1.43	2.02±0.16	14.47±1.63	0.99±0.11	54.01±9.79	2.55±0.18	6.91±1.13	0.11±0.02	20.66±1.53	2.31±0.08
DDPP (+DPP) (ours)	metric	PV	19.51±3.40	2.47±0.28	15.79±1.67	1.07±0.14	43.82±1.82	2.17±0.14	7.33±1.53	0.12±0.02	18.93±2.09	2.11±0.25
DDPP (+DPP) (ours)	metric	BALD	20.54±4.72	2.52±0.34	15.48±1.81	1.03±0.08	43.95±1.68	2.17±0.12	8.01±2.08	0.13±0.03	22.44±4.78	2.67±0.49
DDPP (+DPP) (ours)	metric	SMP	18.45±2.88	2.41±0.26	16.78±3.43	1.14±0.26	43.61±1.61	2.16±0.11	6.92±1.32	0.11±0.02	19.11±2.14	2.16±0.22
DDPP (+OOD) (ours)	-	PV	22.73±7.45	2.65±0.59	19.05±2.95	1.29±0.23	51.11±12.03	2.37±0.34	6.32±0.72	0.10±0.01	16.75±2.31	1.94±0.21
DDPP (+OOD) (ours)	-	BALD	23.85±8.39	2.69±0.58	18.27±3.05	1.22±0.23	52.59±12.08	2.42±0.34	6.59±0.69	0.11±0.01	20.56±3.09	2.50±0.26
DDPP (+OOD) (ours)	-	SMP	22.31±7.80	2.60±0.65	19.86±3.83	1.36±0.29	50.14±9.73	2.32±0.30	6.09±0.67	0.10±0.01	17.76±2.75	2.06±0.23
DDPP (+OOD) (ours)	CER	PV	14.83±1.42	2.05±0.17	14.98±1.36	1.01±0.09	59.14±11.27	2.56±0.24	7.08±1.37	0.11±0.02	19.66±1.25	2.17±0.15
DDPP (+OOD) (ours)	CER	BALD	15.03±1.85	2.08±0.24	14.37±2.22	0.96±0.14	57.48±9.37	2.54±0.26	7.41±1.29	0.12±0.02	25.30±3.36	3.00±0.24
DDPP (+OOD) (ours)	CER	SMP	14.34±1.15	1.99±0.16	15.88±1.96	1.08±0.13	59.32±11.86	2.53±0.20	6.88±1.24	0.11±0.02	21.06±1.96	2.35±0.14
DDPP (+OOD) (ours)	metric	PV	19.03±3.97	2.41±0.34	17.75±5.20	1.10±0.17	48.54±11.38	2.23±0.24	6.92±1.32	0.11±0.02	18.36±1.90	2.05±0.26
DDPP (+OOD) (ours)	metric	BALD	19.33±4.78	2.41±0.40	16.71±7.13	1.02±0.20	49.31±11.87	2.24±0.25	7.21±1.49	0.11±0.02	21.35±4.47	2.54±0.45
DDPP (+OOD) (ours)	metric	SMP	18.55±3.06	2.42±0.27	17.08±3.78	1.14±0.26	43.67±1.77	2.15±0.11	6.71±1.18	0.10±0.02	19.01±2.30	2.16±0.25
SR	CER	MP	14.62±1.62	2.02±0.19	14.56±2.14	1.00±0.14	56.97±9.69	2.53±0.15	6.84±1.41	0.11±0.02	21.31±1.63	2.49±0.25
SR	metric	MP	18.39±2.94	2.40±0.27	16.90±3.12	1.16±0.24	44.54±2.11	2.22±0.15	6.51±1.07	0.10±0.02	20.32±1.68	2.32±0.23
SR (baseline)	-	MP	22.32±8.08	2.58±0.65	17.93±3.84	1.22±0.28	49.48±3.71	2.35±0.25	6.08±0.62	0.10±0.01	18.81±3.35	2.21±0.29

Table 1: Results for methods based on MC dropout and regularization techniques (ELECTRA model). The best results are shown in bold, the best results for each method are underlined.

Results – MD SN

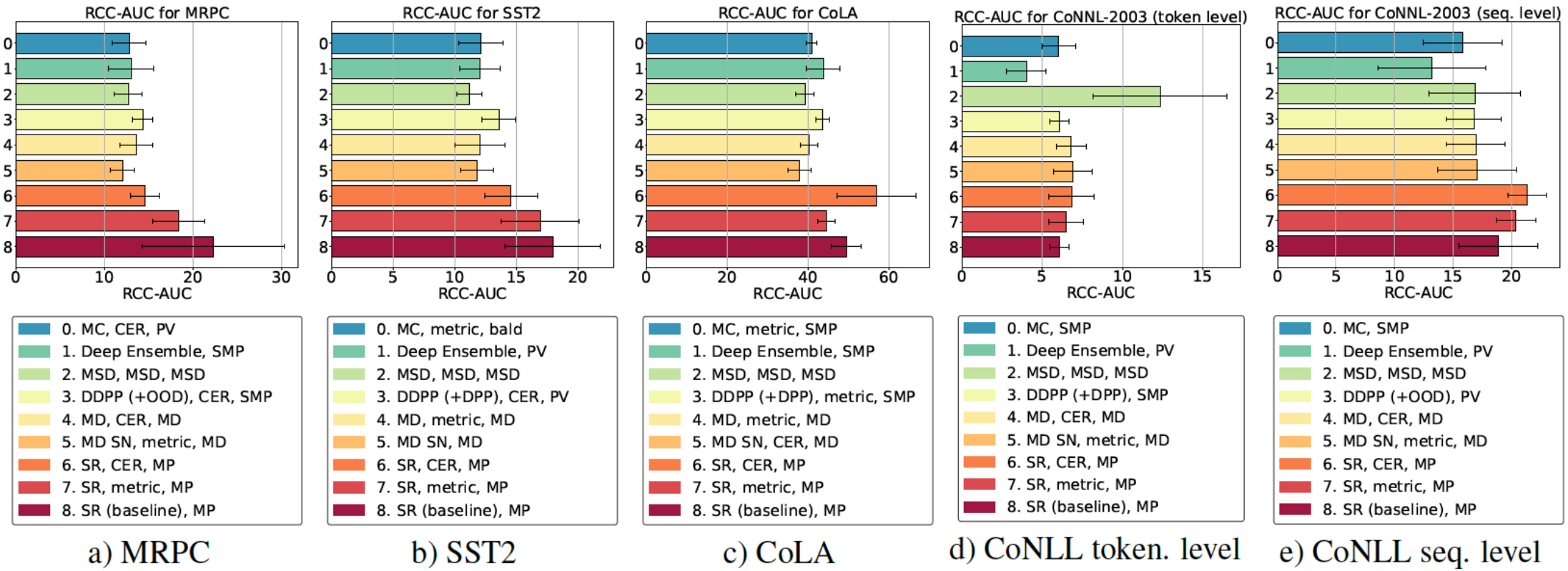
Method	Reg. Type	UE Score	MRPC		SST-2		CoLA		CoNLL-2003 (token level)		CoNLL-2003 (seq. level)	
			RCC-AUC ↓	RPP ↓	RCC-AUC ↓	RPP ↓	RCC-AUC ↓	RPP ↓	RCC-AUC ↓	RPP ↓	RCC-AUC ↓	RPP ↓
MD	-	MD	13.69±1.25	1.88±0.13	13.08±2.58	0.86±0.15	41.73±1.45	1.96±0.04	10.33±3.55	0.15±0.04	17.05±5.07	2.05±0.45
MD	CER	MD	13.61±1.82	1.87±0.22	14.10±2.69	0.96±0.16	42.50±2.65	2.00±0.07	**6.82±0.90**	**0.10±0.01**	**16.92±2.51**	**1.87±0.23**
MD	metric	MD	13.91±2.35	1.89±0.29	12.03±2.04	0.85±0.15	40.29±2.09	2.02±0.09	10.01±2.56	0.15±0.03	17.67±3.92	2.09±0.36
MD SN (ours)	-	MD	13.44±1.28	1.85±0.20	**11.77±1.33**	**0.83±0.08**	40.07±3.62	1.95±0.16	7.21±1.34	0.11±0.02	17.29±3.58	2.01±0.37
MD SN (ours)	CER	MD	14.41±1.96	1.94±0.21	12.32±1.37	0.85±0.10	**37.82±2.91**	**1.90±0.12**	6.95±1.50	0.11±0.02	17.76±4.00	2.06±0.42
MD SN (ours)	metric	MD	**12.04±1.33**	**1.56±0.12**	12.05±1.42	0.84±0.07	39.37±2.00	1.97±0.15	6.90±1.21	0.11±0.02	17.02±3.39	2.01±0.40
SNGP	-	SNGP	14.52±2.48	2.00±0.35	16.08±4.18	1.02±0.18	51.96±1.89	2.64±0.07	56.43±23.03	0.60±0.22	44.80±11.00	5.06±1.01
SR SN	-	MP	18.83±3.89	2.46±0.46	19.02±6.07	1.21±0.35	81.25±12.56	3.40±0.33	7.46±1.39	0.12±0.02	20.13±3.50	2.30±0.26
SR	CER	MP	14.62±1.62	2.02±0.19	14.56±2.14	1.00±0.14	56.97±9.69	2.53±0.15	6.84±1.41	0.11±0.02	21.31±1.63	2.49±0.25
SR	metric	MP	18.39±2.94	2.40±0.27	16.90±3.12	1.16±0.24	44.54±2.11	2.22±0.15	6.51±1.07	0.10±0.02	20.32±1.68	2.32±0.23
SR (baseline)	-	MP	22.32±8.08	2.58±0.65	17.93±3.84	1.22±0.28	49.48±3.71	2.35±0.25	6.08±0.62	0.10±0.01	18.81±3.35	2.21±0.29

Table 2: Results of deterministic methods with different types of regularization (ELECTRA model). The best results are highlighted with the bold font, the best results for each method are underlined.

Results – best results for UE metric

Method	Reg. Type	UE Score	MRPC		SST-2		CoLA		CoNLL-2003 (token level)		CoNLL-2003 (seq. level)	
			RCC-AUC ↓	RPP ↓	RCC-AUC ↓	RPP ↓	RCC-AUC ↓	RPP ↓	RCC-AUC ↓	RPP ↓	RCC-AUC ↓	RPP ↓
MC	-	SMP	14.38±2.07	1.76±0.19	14.00±2.20	0.91±0.15	42.95±5.98	2.01±0.15	6.04±1.03	0.09±0.02	15.79±3.34	1.80±0.35
MC	CER	PV	12.82±1.89	1.60±0.13	12.18±1.20	0.80±0.10	46.84±9.19	2.11±0.23	6.92±1.22	0.10±0.02	17.05±3.14	1.91±0.36
MC	metric	BALD	14.55±2.31	1.73±0.23	12.08±1.79	0.79±0.10	43.76±0.55	2.08±0.07	6.91±1.02	0.10±0.01	17.47±1.85	1.98±0.30
MC	metric	SMP	13.39±1.19	1.72±0.20	13.55±1.65	0.90±0.14	40.88±1.25	2.01±0.09	6.30±0.98	0.10±0.01	16.81±1.40	1.80±0.23
Deep Ensemble	-	PV	20.70±4.24	2.10±0.35	12.02±1.63	**0.71±0.07**	50.15±5.57	2.21±0.19	**4.02±1.24**	**0.06±0.02**	**13.18±4.60**	**1.54±0.57**
Deep Ensemble	-	SMP	13.01±2.57	1.68±0.27	12.13±1.27	0.79±0.08	43.73±4.25	2.05±0.19	4.16±1.37	**0.06±0.02**	13.93±4.88	1.57±0.58
MSD	MSD	DS	12.70±1.61	1.74±0.25	**11.17±1.03**	0.78±0.06	39.21±2.18	**1.90±0.12**	12.34±4.19	0.18±0.05	16.83±3.92	1.94±0.25
DDPP (+DPP) (ours)	-	PV	22.30±7.15	2.58±0.65	16.70±1.38	1.12±0.12	49.75±3.96	2.44±0.29	6.12±0.71	0.10±0.01	16.78±2.44	1.93±0.20
DDPP (+DPP) (ours)	-	SMP	21.79±7.72	2.57±0.68	17.55±3.03	1.19±0.23	47.86±5.51	2.39±0.31	6.08±0.62	0.10±0.01	17.71±2.77	2.05±0.23
DDPP (+DPP) (ours)	CER	PV	15.12±2.27	2.03±0.24	13.56±1.37	0.91±0.14	54.51±8.80	2.58±0.22	6.98±0.98	0.11±0.02	19.44±1.15	2.13±0.17
DDPP (+DPP) (ours)	CER	SMP	14.75±1.43	2.02±0.16	14.47±1.63	0.99±0.11	54.01±9.79	2.55±0.18	6.91±1.13	0.11±0.02	20.66±1.53	2.31±0.08
DDPP (+DPP) (ours)	metric	SMP	18.45±2.88	2.41±0.26	16.78±3.43	1.14±0.26	43.61±1.61	2.16±0.11	6.92±1.32	0.11±0.02	19.11±2.14	2.16±0.22
DDPP (+OOD) (ours)	-	PV	22.73±7.45	2.65±0.59	19.05±2.95	1.29±0.23	51.11±12.03	2.37±0.34	6.32±0.72	0.10±0.01	16.75±2.31	1.94±0.21
DDPP (+OOD) (ours)	-	SMP	22.31±7.80	2.60±0.65	19.86±3.83	1.36±0.29	50.14±9.73	2.32±0.30	6.09±0.67	0.10±0.01	17.76±2.75	2.06±0.23
DDPP (+OOD) (ours)	CER	BALD	15.03±1.85	2.08±0.24	14.37±2.22	0.96±0.14	57.48±9.37	2.54±0.26	7.41±1.29	0.12±0.02	25.30±3.36	3.00±0.24
DDPP (+OOD) (ours)	CER	SMP	14.34±1.15	1.99±0.16	15.88±1.96	1.08±0.13	59.32±11.86	2.53±0.20	6.88±1.24	0.11±0.02	21.06±1.96	2.35±0.14
DDPP (+OOD) (ours)	metric	SMP	18.55±3.06	2.42±0.27	17.08±3.78	1.14±0.26	43.67±1.77	2.15±0.11	6.71±1.18	0.10±0.02	19.01±2.30	2.16±0.25
MD	CER	MD	13.61±1.82	1.87±0.22	14.10±2.69	0.96±0.16	42.50±2.65	2.00±0.07	6.82±0.90	0.10±0.01	16.92±2.51	1.87±0.23
MD	metric	MD	13.91±2.35	1.89±0.29	12.03±2.04	0.85±0.15	40.29±2.09	2.02±0.09	10.01±2.56	0.15±0.03	17.67±3.92	2.09±0.36
MD SN (ours)	-	MD	13.44±1.28	1.85±0.20	11.77±1.33	0.83±0.08	40.07±3.62	1.95±0.16	7.21±1.34	0.11±0.02	17.29±3.58	2.01±0.37
MD SN (ours)	CER	MD	14.41±1.96	1.94±0.21	12.32±1.37	0.85±0.10	**37.82±2.91**	**1.90±0.12**	6.95±1.50	0.11±0.02	17.76±4.00	2.06±0.42
MD SN (ours)	metric	MD	**12.04±1.33**	**1.56±0.12**	12.05±1.42	0.84±0.07	39.37±2.00	1.97±0.15	6.90±1.21	0.11±0.02	17.02±3.39	2.01±0.40
SR	CER	MP	14.62±1.62	2.02±0.19	14.56±2.14	1.00±0.14	56.97±9.69	2.53±0.15	6.84±1.41	0.11±0.02	21.31±1.63	2.49±0.25
SR	metric	MP	18.39±2.94	2.40±0.27	16.90±3.12	1.16±0.24	44.54±2.11	2.22±0.15	6.51±1.07	0.10±0.02	20.32±1.68	2.32±0.23
SR (baseline)	-	MP	22.32±8.08	2.58±0.65	17.93±3.84	1.22±0.28	49.48±3.71	2.35±0.25	6.08±0.62	0.10±0.01	18.81±3.35	2.21±0.29

Results – best results for UE metric



a) MRPC b) SST2 c) CoLA d) CoNLL token. level e) CoNLL seq. level

Figure 1: RCC-AUC↓ of the best UE methods for the ELECTRA model.

Results - ARC

- All rejected instances by certain uncertainty level are labeled with an oracle (GT labels)

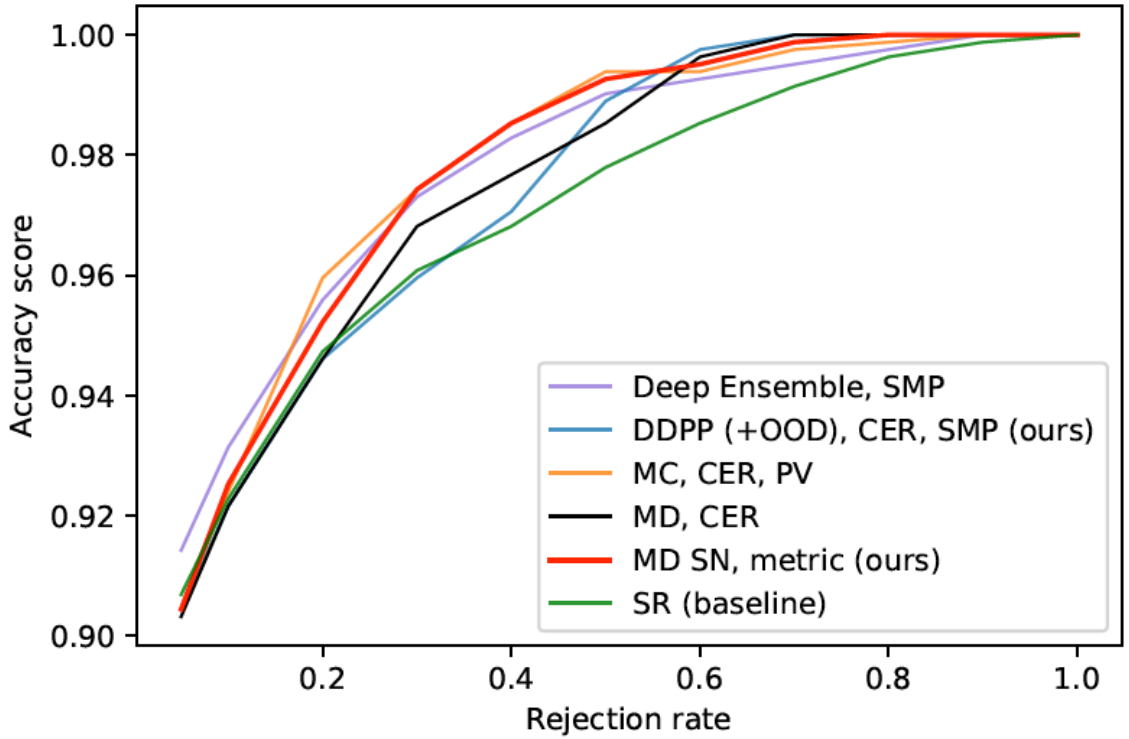


Figure 2: Median values of accuracy rejection curves for selected methods on MRPC (ELECTRA model).

Conclusion

- Two computationally cheap UE methods: DDPP + DPP/OOD; MD+SN
- Extensive experiments to investigate UE on two tasks
- Methods on par with MC dropout, and outperform SR baseline.

UE on WSD?

- Aleatoric uncertainty: disagreement; multi-label; fuzziness
- Epistemic uncertainty: MFS bias; OOD sense

- Less work on this topic (one in 2009 [Zhu et al., 2009])

References

- Zhu, Jingbo, et al. "Active learning with sampling by uncertainty and density for word sense disambiguation and text classification." *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. 2008.
- Xin, Ji, et al. "The art of abstention: Selective prediction and error regularization for natural language processing." *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)*. 2021.
- Zhang, Xuchao, et al. "Mitigating Uncertainty in Document Classification." *Proceedings of NAACL-HLT*. 2019.
- Lee, Kimin, et al. "A simple unified framework for detecting out-of-distribution samples and adversarial attacks." *Advances in neural information processing systems* 31 (2018).
- Shelmanov, Artem, et al. "How Certain is Your Transformer?." *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2021.

清華大學
Tsinghua University

Q & A

THANK YOU

SOTA models

Model

Name	Task	Author	Conference
Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders [CODE] [Paper] [Video] (BEM, ACL'20)	Token classification	Terra Blevins and Luke Zettlemoyer (Facebook ai)	ACL'20
Framing Word Sense Disambiguation as a Multi-Label Problem for Model-Agnostic Knowledge Integration [CODE] [Paper]		Simone Conia, Roberto Navigli	EACL'21
Breaking through the 80% glass ceiling: Raising the state of the art in Word Sense Disambiguation by incorporating knowledge graph information [paper][CODE][Video] (EWISER, ACL'20)	Token classification	Michele Bevilacqua and Roberto Navigli	ACL'20
EViLBERT: Learning Task-Agnostic Multimodal Sense Embeddings [paper][Short Video][Long Video] [Website&CODE] (EViLBERT, IJCAI'20)		Agostina Calabrese , Michele Bevilacqua and Roberto Navigli	IJCAI'20
Zero-shot Word Sense Disambiguation using Sense Definition Embeddings [Video+Code+Pdf] (EWISE, ACL'19)		Sawan Kumar, Sharmistha Jat, Karan Saxena, Partha Talukdar	ACL'19
Sense Vocabulary Compression through the Semantic Knowledge of WordNet for Neural Word Sense Disambiguation [PDF+CODE] (SVC)	Token classification	Loïc Vial, Benjamin Lecouteux, Didier Schwab	GWNC'19*
Improved Word Sense Disambiguation Using Pre-Trained Contextualized Word Representations Christian [PDF+CODE] (GLU)		Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan	EMNLP-IJCNLP 2019
GlossBERT: BERT for Word Sense Disambiguation with gloss knowledge	Token classification	Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang	EMNLP-IJCNLP 2019

*GWNC: Proceedings of the 10th Global Wordnet Conference

WSD datasets

- Evaluation: the Senseval/SemEval competitions: held from 1998
- (Training) corpus: SemCor (manually) & OMSTI (automatically)

	#Docs	#Sents	#Tokens	#Annotations	#Sense types	#Word types	Ambiguity
Senseval-2	3	242	5,766	2,282	1,335	1,093	5.4
Senseval-3	3	352	5,541	1,850	1,167	977	6.8
SemEval-07	3	135	3,201	455	375	330	8.5
SemEval-13	13	306	8,391	1,644	827	751	4.9
SemEval-15	4	138	2,604	1,022	659	512	5.5
SemCor	352	37,176	802,443	226,036	33,362	22,436	6.8
OMSTI	-	813,798	30,441,386	911,134	3,730	1,149	8.9

#Sense types: number of unique sense (sense vocabulary)
Ambiguity: number of candidate senses on average

Dataset statistics (P2)

Datasets	Train	Test	# Labels
MRPC	3.7K	0.4K	2
CoLA	8.6K	1.0K	2
SST-2	67.3K	0.9K	2
CoNLL-2003	14.0K/203.6K	3.5K/46.4K	9

Table 4: Dataset statistics. The table presents the number of sequences for the training and test parts of the datasets. For CoNLL-2003, the table presents both the number of sequences and tokens because for NER, we evaluate both sequence-level and token-level UE scores. For the datasets from the GLUE benchmark (MRPC, CoLA, SST-2), we used the available validation set as the test set.