A Closer Look at Probability Calibration of Knowledge Graph Embedding

Ruiqi Zhu[1], Fangrong Wang[1], Alan Bundy[1], Xue Li[1], Kwabena Nuamah[1], Lei Xu[2], Stefano Mauceri[2], Jeff Z. Pan[1,3*]
[1]School of Informatics, the University of Edinburgh, UK; [2] Ireland Research Centre, Huawei
[3]Edinburgh Research Centre, CSI, Huawei

ABSTRACT

When the estimated probabilities do not match the relative frequencies, we say these estimated probabilities are *uncalibrated* [39], which may cause incorrect decision making, and is particularly undesired in high-stakes tasks [45]. Knowledge Graph embedding models are reported to produce uncalibrated probabilities [36], e.g., for all the triples predicted with probability 0.9, the percentage of them being truly correct triples is not 90%. In this article, we take a closer look at this problem. First, we confirmed the issue that typical KG Embedding models are uncalibrated. Then, we show how off-the-shelf calibration techniques can be used to mitigate this issue, among which binning-based calibration produces more calibrated probabilities. We also investigated the possible reasons for the uncalibrated probabilities and found that the *expit transform*, the way used to convert embedding scores into probabilities, is ineffective in most cases.

KEYWORDS

Knowledge Graph Embedding, Probability Calibration

ACM Reference Format:
Ruiqi Zhu[1], Fangrong Wang[1], Alan Bundy[1], Xue Li[1], Kwabena Nuamah[1], Lei Xu[2], Stefano Mauceri[2], Jeff Z. Pan[1,3]. 2022. A Closer Look at Probability Calibration of Knowledge Graph Embedding. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Knowledge Graphs (KG) [22] are becoming popular and gaining increasing usage in various application scenarios. Probabilistic Knowledge Graphs (PKG), in which each triple is assigned a probability of the triple being correct, play an important role in scenarios of uncertainty [9, 31], e.g., drug discovery [44].

One approach to assigning probabilities to triples is to train embedding models [4], e.g., TransE [3] or ComplEx [37], for knowledge graphs, and then use the scoring function of the trained embedding model to score the new triples:

$$score = f_{embed}(\langle \vec{s}, \vec{p}, \vec{o} \rangle)$$

where f_{embed} is the scoring function of the embedding model, and s, p, o represent subject, predicate and object, respectively. Prior work suggested that these scores can be converted into probabilities via *expit transform* [19, 36], i.e., passing these scores through the sigmoid function as follows.

$$prob = \sigma(score) = \frac{1}{(1 + exp(-score))}$$

Later work [36] showed that the probabilities obtained in this way are uncalibrated; e.g., for all the triples with probability 0.9, the percentage of them being correct triples w.r.t. the real world is not 90%. Thus, these expit-transformed probabilities need to be calibrated

$$prob^* = f_{calib}(prob)$$

where f_{calib} is a calibration model, and $prob^*$ are the calibrated probabilities that do not over-estimate or under-estimate the truth of triples.

We looked closer at the research of probability calibration for knowledge graph embedding, with the following contributions:

(1) We stressed [1] that not all expit-transformed scores are appropriate to be interpreted as probabilities. Also, we argue that probability calibration can serve as a more accurate technique to convert embedding model scores into probabilities.
(2) Though expit-transformed scores of some embedding models can be interpreted as probabilities, we found that these probabilities are uncalibrated, and thus calibration is needed.
(3) We provide empirical evidence for a useful rule of thumb [21] for how to choose calibration techniques: for a large set of held-out data (say, over 10 thousand triples), binning-based calibration techniques perform better, such as Isotonic Regression and Histogram Binning. Otherwise, scaling-based techniques, such as Platt Scaling, are more suitable.

2 PRELIMINARIES

In this section, we briefly explain some important notions used in our work.

Knowledge Graphs [22] are represented in a standard format for graph-structured data such as RDF. A *knowledge graph* $\mathcal{G}$ is a tuple $(\mathcal{E}, \mathcal{R}, \mathcal{T})$, where $\mathcal{E}$ is a set of entities, $\mathcal{R}$ is a set of relation types, and $\mathcal{T}$ is a set of relational triple $\langle s, p, o \rangle$, where $s, o \in \mathcal{E}$ are

*Correspondence author: Jeff Z. Pan, https://knowledge-representation.org/j.z.pan/

[1]We are not the first to show this phenomenon, but unfortunately still many people mess up.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

respectively the *head* and *tail* entities of the triple, and $p \in \mathcal{R}$ is the *edge* of the triple connecting head and tail [23].

Knowledge Graph Embedding is a family of algorithms to map the entities and relations of a knowledge graph to a m-dimension vector space $\mathbb{R}^m$. A KG embedding model usually defines a scoring function $f(\langle \vec{s}, \vec{p}, \vec{o} \rangle)$ that evaluates the truth/correctness of a triple, where $\vec{s}, \vec{p}, \vec{o} \in \mathbb{R}^m$ are the relevant embeddings of s, p, o. The model then strives to find the best embedding for all entities and relations, such that the positive (correct) triples get as high scores as possible while the negative (incorrect) triples get as low scores as possible.

Probability Calibration is the technique to adjust the uncalibrated probabilities, or directly transform classifier scores of no probability meanings into probabilities that satisfy probability axioms and have probability semantics.

Formally, consider binary classification tasks. Given a set of samples $(X, y) \in \mathcal{D}$, if $\forall \beta \in [0,1]$, we have $fr\big(X|pr(X) = \beta\big) = \beta$, where $fr(X)$ represents the frequency of X being a positive sample, and $pr(X)$ represents the predicted probability of X being a positive sample, we say the predicted probabilities $pr(X)$ are calibrated. Otherwise, we say they are uncalibrated.

Calibrated probabilities are desired, especially in high-stake decision-making tasks, like medical diagnosis, autonomous driving, etc. Uncalibrated probabilistic models will lead to under-estimated or over-estimated risks [11, 38], while calibrated probabilities are necessary to make optical decisions [15, 45]. Zhao et al [45] mathematically formalised the benefits of calibrated probabilities as **No Regret Decision Making** and **Accurate Loss Estimation**.

To evaluate how well a set of probabilities are calibrated, metrics such as Brier Score, Negative Log Loss, and Expected Calibration Error are available [18]. They are defined as:

$$BS = \frac{1}{N} \sum_{i=1}^{N} (p_i - y_i)^2$$

$$NLL = -\frac{1}{N} \sum_{i=1}^{N} y_i log(p_i) + (1 - y_i) log(1 - p_i)$$

where N is the number of samples, $p \in [0,1]$ is the predicted probability of the ith sample, and $y \in \{0,1\}$ is the relevant truth label.

$$ECE = \frac{1}{b} \sum_{j}^{b} |pr_j - fr_j|$$

where b is the number of bins[2] for the unit interval, pr_j and fr_j is the average probability and relative frequency of the samples grouped in the jth bin.

3 RELATED WORKS

As the concept of knowledge graph was popularised by Google in 2012 [22], in 2013 Bordes et al., had proposed TransE [3], a forerunner of KG Embedding models. Afterwards, subsequent new KG embedding models were proposed. Just to name a few typical ones, Tabacof's experiment used ComplEx [37], DistMult [42], and HoLE [20].

Two widely used probability calibration techniques are Platt Scaling (or Logistic Calibration) [26] and Isotonic Regression [21]. There are many more calibration techniques, such as Beta Calibration [14] and Histogram Binning [43]. As deep learning progressed rapidly, people discovered that probabilistic outputs of deep neural networks, particularly those with Batch Norm layers, were uncalibrated [10], and proposed new calibration techniques for modern deep neural networks, e.g., Temperature Scaling [10]. Broadly speaking, Beta Calibration and Temperature Scaling are variants of Platt Scaling and we call them *scaling-based techniques*. While Histogram Binning is a variant of Isotonic Regression, we call them *binning-based techniques*. According to our evaluation, the binning-based techniques perform better in large datasets.

Although KG embedding and Probability Calibration are being actively studied, calibrating KG embedding models is relatively under-explored. To the best of our knowledge, Tabacof et al [36] were the first to look at this problem. They reported the uncalibrated nature of KG embedding models and used calibrated probabilities to perform the triple classification task. To follow up, Pezeshkpour et al [25] showed that different negative sampling strategies can have different effects on the calibration. Safavi et al [30] then used calibration to improve the trustworthiness of link prediction results, which is a main downstream application of KG embedding. Besides, Rao [29] investigated calibrating Knowledge Graph under the closed-world assumption and open-world assumption. Indeed, these are all the recent works we found about probability calibration for knowledge graph embedding.

Building on top of the prior works, we conducted extended experiments to test several calibration techniques on several datasets related to the problem of KG embedding. We noted that prior works [19, 36] mistakenly apply expit transforms to obtain probabilities to measure the correctness of a given triple, resulting in bad probabilities that are uncalibrated. We suggested calibration as a better approach than expit transform.

4 EXPIT-TRANSFORMED SCORES AS PROBABILITIES?

Depending on the scoring function of KG embedding models, expit-transformed scores sometimes can be interpreted as probabilities but sometimes not. We are not the first to point out this issue. It has even been noted in some libraries documentation[3]. For instance, TransE adopts such a distance-based scoring function:

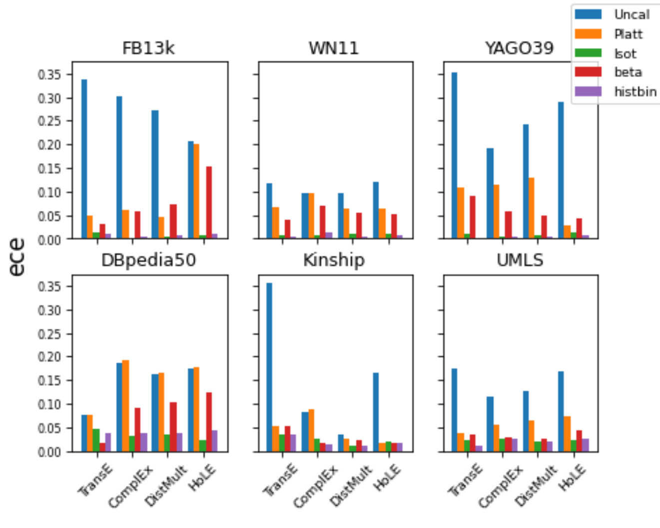
$$f_{TransE}(\langle \vec{s}, \vec{p}, \vec{o} \rangle) = -||\vec{s} + \vec{p} - \vec{o}||_2$$

Hence $f_{TransE}(\langle \vec{s}, \vec{p}, \vec{o} \rangle) \in [-\infty, 0]$, and thus $\sigma\big(f_{TransE}(s, p, o)\big) \in [0, 0.5]$. That is to say, the expit-transformed scores of TransE are always lower than 0.5, which can hardly be recognised as probabilities, regardless of the truth of a triple. Any embedding models adopting distance-based scoring functions as TransE, such as TransD [12], TransR [41], TransH [17], RotatE [35], PairRE [6] and BoxE [1] will suffer from this problem.

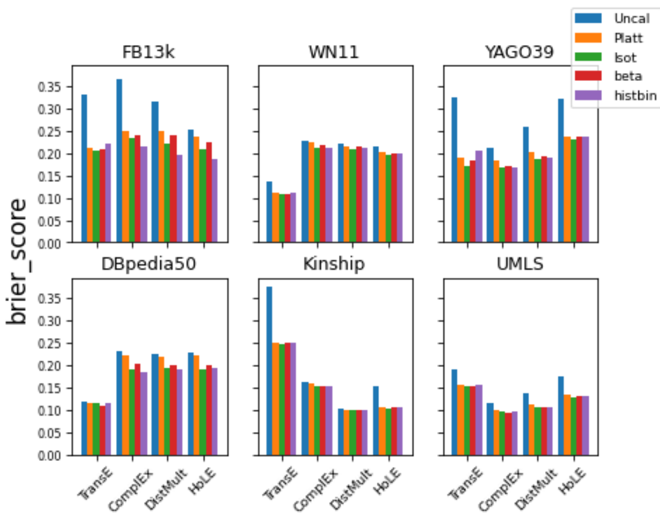
Some may suggest it is not a problem because we can always map the scale to the unit interval, for example, doubling the scale of expit-transformed scores of TransE so that now the range turns

[2] We group triples according to their estimated probabilities, e.g., all the triples whose probabilities within $[0.1, 0.2]$ are grouped in one bin.

[3] https://pykeen.readthedocs.io/en/stable/reference/models.html Accessed on October 2, 2022



(a) Expected Calibration Error



(b) Brier Score

Figure 1: Bar charts of ECE and BS for the probabilities produced by expit transform and the probabilities produced by various calibration techniques per model per dataset. The smaller ECE or BS, the better calibrated.

from [0, 0.5] to [0, 1], and obey the probability axioms [40]. In our later experiments (§5.1), the expit-transformed values of TransE did achieve relatively high accuracy in the triple classification task. Nevertheless, it is not the case when it turns to other embedding models. As shown in Figure 2, the doubled expit-transformed values of TransR and RotatE[4] are still lower than 0.5.

Whether the expit-transformed scores are probabilities could be arguable, but in the following experiments, we can show that even if we consider them as probabilities, they are uncalibrated, and thus cannot be used in high stake applications.

[4]These two models are not implemented in Ampligraph, so we used the PyKEEN [2] library implementations.

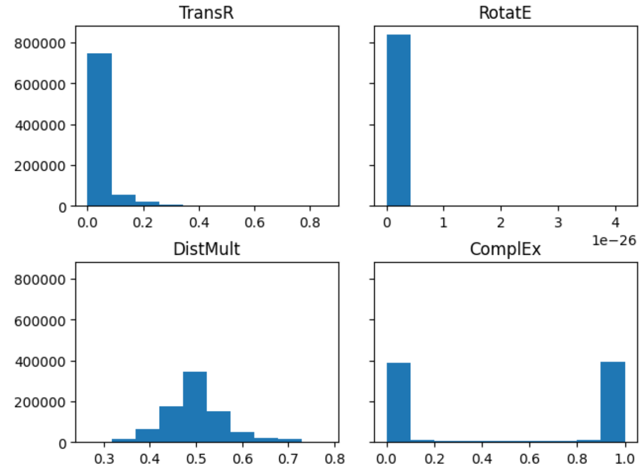


Figure 2: Histograms of doubled expit-transformed values of TransR, and RotatE, compared with DistMult and ComplEx (not doubled). Models were trained on UMLS dataset, optimising the NLL loss, with 500 epochs and early-stopping trick.

5 EXPERIMENT AND RESULTS

We conducted experiments[5] to examine the following hypothesis:

(1) Expit-transformed probabilities of current KG Embedding Models are uncalibrated, but off-the-shelf calibration techniques can effectively make the uncalibrated probabilities calibrated, producing more accurate probability estimations (see §5.1).
(2) Binning-based techniques (Isotonic Regression and Histogram Binning) generally work better than scaling-based ones (Platt Scaling and Beta Calibration) when large datasets are available (see §5.2).

Extending the setting of the previous work by Tabacof et al [36], in our experiment, we trained 4 typical KG embedding models, TransE [3], ComplEx [37], DistMult [42], and HoLE [20] on 6 datasets: FB13k [33], WN11 [33], YAGO39 [8], DBpedia50 [32], Kinship [13], and UMLS [13]. Each dataset is split into 3 subsets for training, calibration, and testing. The calibration and testing sets of FB13, WN11 and YAGO39 have ground truth negative samples, while the other 4 don't. Therefore, we generated synthetic negative samples via the corruption and local closed world assumption. In all datasets, we have balanced positive and negative samples.

We used the implementation of Knowledge Graph Embedding Models from AmpliGraph[6] [7] and the implementation of calibration techniques from NetCal[7] [16]. We trained each model for 500 epochs to optimise the Negative Log Loss, using early-stopping to avoid over-fitting. The vector dimensionality is set to 100. We used the Adam optimiser with an initial learning rate of $1e-4$.

[5]Code will be available at https://github.com/TREAT-UOE/kgcal
[6]https://github.com/Accenture/AmpliGraph visited on October 2, 2022
[7]https://github.com/fabiankueppers/calibration-framework visited on October 2, 2022

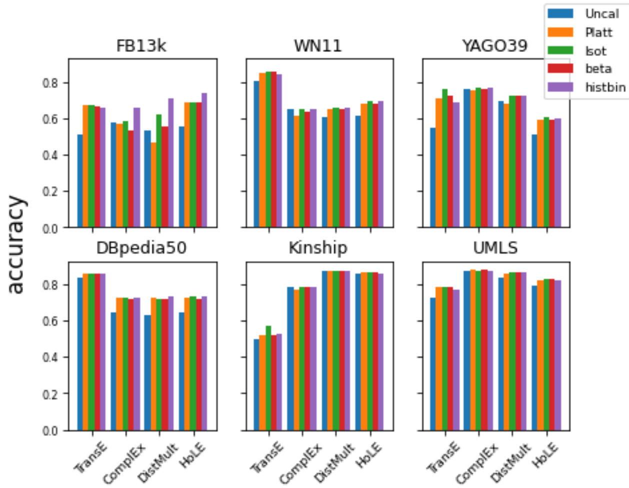


Figure 3: Accuracy of triple classification task in different datasets, using probabilities and a natural threshold $\tau = 0.5$

5.1 Uncalibrated Probabilities

To evaluate hypothesis (1), our goal is to compare the expit-transformed probabilities and calibrated probabilities and show whether the former incur higher calibration errors.Firstly, we trained KG embedding models on a training set ($Train_E$) and computed the expit-transformed probabilities of triples in the test set. Specifically, we doubled the expit-transformed values of TransE so that the range of them is turned from $[0, 0.5]$ to $[0, 1]$. Then, we trained a calibration model on a held-out set ($Train_C$) and obtained the calibrated probabilities of triples in the testing set via the calibration model. We compared the expit-transformed probabilities and the calibrated probabilities in Figure 1, which illustrates that expit-transformed values get higher ECEs and BSs than calibrated ones, meaning that the KG embedding models are more or less uncalibrated, and almost all calibration techniques produced better-calibrated probabilities than the expit-transformed ones.

We use these probabilities to perform the triple classification task with 0.5 as the threshold. We chose 0.5 because it is the natural threshold of probabilities. Without further elaboration, we tend to believe that a statement with a probability higher than 0.5 is likely to be true, while a statement with a probability lower than 0.5 is likely to be false. Figure 3 shows that the calibrated probabilities can serve as a better indicator to classify the positive triples from the negative ones than the uncalibrated ones. In most cases, calibrated probabilities can do at least as good as uncalibrated probabilities. In some cases, calibrated probabilities can significantly lift the classification accuracy. We also noted that the expit-transformed probabilities of TransE (doubled) in some datasets achieve closed accuracy as the corresponding calibrated probabilities, which means in classification tasks they can serve as probabilities, but no better than calibrated ones.

These results suggest that calibration is a better way than expit transform to convert embedding scores into more calibrated and accurate probabilities. Expit-transformed probabilities, after the range adapted to [0,1], should be used only when no extra data (the calibration set $Train_C$) is available to train a calibration model.

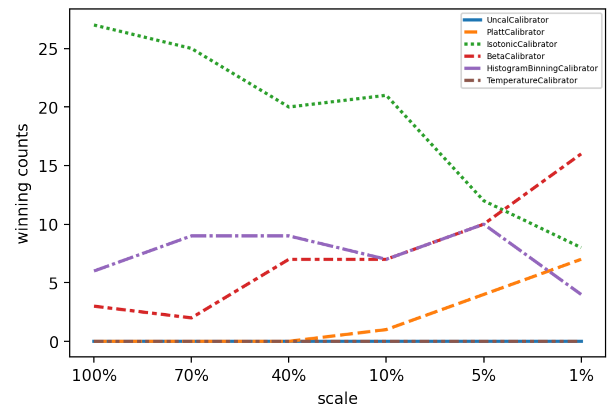


Figure 4: Number of winning counts for different calibration techniques for the 4 KG embedding models when the calibration sets of FB13, WN11, and YAGO 39 shrink. For each calibration result, we compute all the 3 metrics (BS, NLL, and ECE), so that every bin in the figure has 36 counts in total.

5.2 Binning-based Calibration

During the experiment, we observed that binning-based calibration (Isotonic and Histogram) performs better in general. We also noticed that binning-based methods dominated in FB13k, WN11 and YAGO39, which has more data than the rest. Previous work also suggested that binning-based methods tend to overfit, especially on smaller datasets [21]. Thus, to evaluate hypothesis (2), we took these 3 datasets, and gradually shrink the size of the calibration sets by randomly sampling k% of them, and compare the number of wins in terms of BS, NLL, and ECE between binning-based and scaling-based methods. We plotted the results in Figure 4.

Results show that the performance of binning-based calibration techniques dominates at the beginning. As the size of the calibration sets shrinks, the winning count of Isotonic Regression and Histogram Binning decreases, while that of Platt Scaling and Beta Calibration increases. This implies that we should prefer binning-based calibration when large datasets are available (e.g. over 10k triples). When the dataset is relatively small, determining which calibration technique is better requires careful empirical evaluation.

6 CONCLUSION

We stressed that not all expit-transformed scores are appropriate to be interpreted as probabilities. What is worse, probabilities obtained by expit transform are generally uncalibrated for various KG embedding scores on various datasets. However, off-the-shelf calibration techniques can effectively calibrate these probabilities. If large datasets (over 10k triples) are available, binning-based techniques, including Isotonic Regression and Histogram Binning produced the best calibrated probabilities. In a long run, we will still need to compare the usefulness of probability against other kinds of uncertainties, like possibility [27, 28] and fuzziness [24, 34]. What's more, in this research we only focusd on those widely used embedding models. In the future, we will look at the recently proposed models, like DualE [5] and JointE [46].

ACKNOWLEDGMENTS

The authors would like to thank Huawei for supporting the research on which this paper was based under grant CIENG4721/LSC. We also acknowledge the support of UKRI grant EP/V026607/1, ELIAI (The Edinburgh Laboratory for Integrated Artificial Intelligence) and EPSRC grant no EP/W002876/1. Additional appreciation to the anonymous reviewers who offered useful comments for improving the quality of the paper.

REFERENCES

[1] Ralph Abboud, Ismail Ceylan, Thomas Lukasiewicz, and Tommaso Salvatori. 2020. Boxe: A box embedding model for knowledge base completion. *Advances in Neural Information Processing Systems* 33 (2020), 9649–9661.

[2] Mehdi Ali, Max Berrendorf, Charles Tapley Hoyt, Laurent Vermue, Sahand Sharifzadeh, Volker Tresp, and Jens Lehmann. 2021. PyKEEN 1.0: A Python Library for Training and Evaluating Knowledge Graph Embeddings. *Journal of Machine Learning Research* 22, 82 (2021), 1–6. http://jmlr.org/papers/v22/20-825.html

[3] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in neural information processing systems* 26 (2013).

[4] Hongyun Cai, Vincent W Zheng, and Kevin Chen-Chuan Chang. 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering* 30, 9 (2018), 1616–1637.

[5] Zongsheng Cao, Qianqian Xu, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. 2021. Dual quaternion knowledge graph embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6894–6902.

[6] Linlin Chao, Jianshan He, Taifeng Wang, and Wei Chu. 2021. PairRE: Knowledge Graph Embeddings via Paired Relation Vectors. (Aug. 2021), 4360–4369. https://doi.org/10.18653/v1/2021.acl-long.336

[7] Luca Costabello, Sumit Pai, Chan Le Van, Rory McGrath, Nicholas McCarthy, and Pedro Tabacof. 2019. AmpliGraph: a Library for Representation Learning on Knowledge Graphs. https://doi.org/10.5281/zenodo.2595043

[8] Tim Dettmers, Pasquale Minervini, Pontus Stenetorp, and Sebastian Riedel. 2018. Convolutional 2d knowledge graph embeddings. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[9] Jhonatan Garcia, Jeff Z. Pan, Achille Fokoue, Katia Sycara, Yuqing Tang, and Federico Cerutti. 2015. Handling uncertainty: An extension of DL-Lite with Subjective Logic. In *Proc. of 28th International Workshop on Description Logics (DL 2015)*.

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.

[11] Lars Holmberg and Andrew Vickers. 2013. Evaluation of prediction models for decision-making: beyond calibration and discrimination. *PLoS medicine* 10, 7 (2013), e1001491.

[12] Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*. 687–696.

[13] Charles Kemp, Joshua B Tenenbaum, Thomas L Griffiths, Takeshi Yamada, and Naonori Ueda. 2006. Learning systems of concepts with an infinite relational model. In *AAAI*, Vol. 3. 5.

[14] Meelis Kull, Telmo Silva Filho, and Peter Flach. 2017. Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*. PMLR, 623–631.

[15] Meelis Kull, Telmo M Silva Filho, and Peter Flach. 2017. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics* 11, 2 (2017), 5052–5080.

[16] Fabian Kuppers, Jan Kronenberger, Amirhossein Shantia, and Anselm Haselhoff. 2020. Multivariate confidence calibration for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 326–327.

[17] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Twenty-ninth AAAI conference on artificial intelligence*.

[18] Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

[19] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2015. A review of relational machine learning for knowledge graphs. *Proc. IEEE* 104, 1 (2015), 11–33.

[20] Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. 2016. Holographic embeddings of knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[21] Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*. 625–632.

[22] J.Z. Pan, G. Vetere, J.M. Gomez-Perez, and H. Wu (Eds.). 2017. *Exploiting Linked Data and Knowledge Graphs for Large Organisations*. Springer.

[23] Jeff Z. Pan. 2009. Resource Description Framework. In *Handbook on Ontologies*. 71–90. https://doi.org/10.1007/978-3-540-92673-3_3

[24] Jeff Z. Pan, Giorgos Stamou, Vassilis Tzouvaras, and Ian Horrocks. 2005. f-SWRL: A Fuzzy Extension of SWRL. In *Proc. of the International Conference on Artificial Neural Networks (ICANN 2005)*.

[25] Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2020. Revisiting evaluation of knowledge base completion models. In *Automated Knowledge Base Construction*.

[26] John Platt et al. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers* 10, 3 (1999), 61–74.

[27] Guilin Qi, Jeff Z. Pan, and Qiu Ji. 2007. A Possibilistic Extension of Description Logics. In *Proc. of 2007 International Workshop on Description Logics (DL2007)*.

[28] Guilin Qi, Jeff Z. Pan, and Qiu Ji. 2007. Extending Description Logics with Uncertainty Reasoning in Possibilistic Logic. In *the Proc. of the 9th European Conference on Symbolic and Quantitave Approaches to Reasoning with Uncertainty (ECSQARU'2007)*. 828–839.

[29] Aishwarya Rao. 2021. *Calibrating Knowledge Graphs*. Rochester Institute of Technology.

[30] Tara Safavi, Danai Koutra, and Edgar Meij. 2020. Evaluating the Calibration of Knowledge Graph Embeddings for Trustworthy Link Prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8308–8321.

[31] Murat Sensoy, Jeff Z. Pan, Achille Fokoue, Mudhakar Srivatsa, and Felipe Meneguzzi. 2012. Using Subjective Logic to Handle Uncertainty and Conflicts. In *Proc. of the 2012 International Symposium on Advances in Trusted and Secure Information Systems (TrustCom 2012)*.

[32] Baoxu Shi and Tim Weninger. 2018. Open-world knowledge graph completion. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[33] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. *Advances in neural information processing systems* 26 (2013).

[34] Giorgos Stoilos, Giorgos B. Stamou, and Jeff Z. Pan. 2006. Handling Imprecise Knowledge with Fuzzy Description Logic. In *Proceedings of the 2006 International Workshop on Description Logics (DL2006)*.

[35] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. Rotate: Knowledge graph embedding by relational rotation in complex space. *arXiv preprint arXiv:1902.10197* (2019).

[36] Pedro Tabacof and Luca Costabello. 2020. Probability Calibration for Knowledge Graph Embedding Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=S1g8K1BFwS

[37] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International conference on machine learning*. PMLR, 2071–2080.

[38] Ben Van Calster and Andrew J Vickers. 2015. Calibration of risk prediction models: impact on decision-analytic performance. *Medical decision making* 35, 2 (2015), 162–169.

[39] Bas C Van Fraassen. 1983. Calibration: A frequency justification for personal probability. In *Physics, philosophy and psychoanalysis*. Springer, 295–319.

[40] Susan Vineberg. 2016. Dutch Book Arguments. In *The Stanford Encyclopedia of Philosophy* (Spring 2016 ed.), Edward N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[41] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 28.

[42] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575* (2014).

[43] Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Icml*, Vol. 1. Citeseer, 609–616.

[44] Xiangxiang Zeng, Xinqi Tu, Yuansheng Liu, Xiangzheng Fu, and Yansen Su. 2022. Toward better drug discovery with knowledge graph. *Current opinion in structural biology* 72 (2022), 114–126.

[45] Shengjia Zhao, Michael Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. 2021. Calibrating predictions to decisions: A novel approach to multi-class calibration. *Advances in Neural Information Processing Systems* 34 (2021), 22313–22324.

[46] Zhehui Zhou, Can Wang, Yan Feng, and Defang Chen. 2022. JointE: Jointly utilizing 1D and 2D convolution for knowledge graph embedding. *Knowledge-Based Systems* 240 (2022), 108100. https://doi.org/10.1016/j.knosys.2021.108100