CLOSED BOUNDARY LEARNING FOR NLP CLASSIFICATION TASKS WITH THE UNIVERSUM CLASS

Anonymous authors
Paper under double-blind review

ABSTRACT

The Universum class, often known as the *other* class or the *miscellaneous* class, is defined as a collection of samples that do not belong to any class of interest. It is a typical class that exists in many classification-based tasks in natural language processing (NLP), such as relation extraction, named entity recognition, sentiment analysis, etc. During data labeling, a significant number of samples are annotated as Universum because there are always some samples that exist in the dataset but do not belong to preset target classes and are not of interest in the task. The Universum class exhibits very different properties, namely **heterogeneity** and **lack of representativeness in training data**; however, existing methods often treat the Universum class equally with the classes of interest. Although the Universum class only contains uninterested samples, improper treatment will result in the misclassification of samples of interest. In this work, we propose a closed boundary learning method that treats the Universum class and classes of interest differently. We apply closed decision boundaries to classes of interest and designate the area outside all closed boundaries in the feature space as the space of the Universum class. Specifically, we formulate the closed boundaries as arbitrary shapes, propose a strategy to estimate the probability of the Universum class according to its unique property rather than the within-class sample distribution, and propose a boundary learning loss to learn decision boundaries based on the balance of misclassified samples inside and outside the boundary. We evaluate our method on 6 state-of-the-art works in 3 different tasks, and the performance of all 6 works is improved. Our code will be released on GitHub.

1 INTRODUCTION

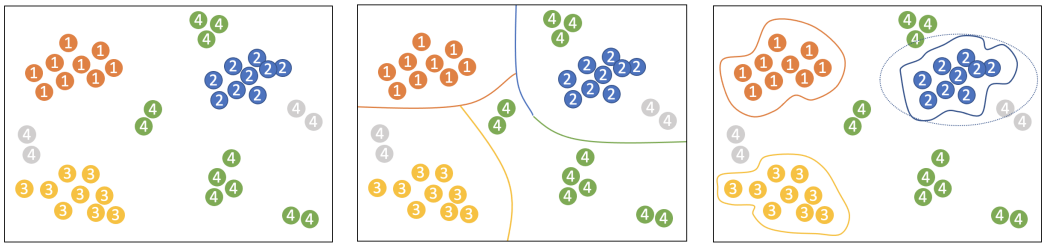
In classification-based tasks of NLP, quite often we encounter a class named as *other* class, *miscellaneous* class, *neutral* class or *outside (O)* class. Such a class is a collection of samples that do not belong to any class of interest, such as samples of *no relation* class in relation extraction task. We adopt the terminology in (Weston et al., 2006) to designate all such classes as the *Universum class* (U). Universum class exits in various classification-based problems in NLP, such as relation extraction (RE) (Zhang et al., 2017), named entity recognition (NER) (Tjong Kim Sang & De Meulder, 2003), sentiment analysis (SA) (Tjong Kim Sang & De Meulder, 2003), and natural language inference (NLI) (Bowman et al., 2015). To distinguish the Universum class and the rest of the classes, we call the classes of interest as *target classes* (T). The set of all classes (A) in training and testing data can be expressed as $A = U \cup T$

- *Universum class*: A collection of samples that do not belong to any class of interest.
- *Target class*: A class of interest in the task, i.e., one of the classes other than the Universum class.

The sample compositions of the Universum class and target classes are usually very different. Figure 1(a) provides some samples of a target class (*entity-destination*) and the Universum class (*other*) in relation extraction. From the examples, we can observe that the *entity-destination* samples adhere to an intrinsic pattern: an entity goes somewhere. However, the three examples of the *other* relation type are vastly dissimilar and do not exhibit any common intrinsic pattern. In fact, the Universum samples are labeled according to an extrinsic pattern: they do not belong to any of the predefined target classes.

Entity-Destination Relation	*Other* Relation
The famous actress arrived at the airport.	The captain and crews are grateful for the support.
Quake survivors moved into makeshift houses.	The room was filled with huge canvases.
The research team is going into the deep jungle.	The stories are narrated through dance.
...	...

(a)



(b) (c) (d)

Figure 1: Illustration of distinction between the Universum class and target classes and demonstration of open boundary and closed boundary classifiers for tasks with the Universum class. (a) Samples selected from the SemEval 2010 Task 8 dataset on relation extraction. (b) The distribution of target classes (class 1, 2, 3) and the Universum class (class 4). In particular, the gray samples represent Universum samples in the test set that are not represented by the training data. (c) The open decision boundaries obtained by traditional classifiers. (d) The arbitrary closed boundaries obtained by our proposed method.

We further highlight the differences between the Universum class and target classes in two properties.
(1) **Heterogeneity**: The universum class is composed of heterogeneous samples, which may form multiple clusters in the feature space, as illustrated by the green samples in Figure 1(b). This is because the Universum class, as the class name "other" implies, contains all potential implicit classes that are not explicitly defined in the task. For example, in the *other* samples given in Figure 1(a), implicit classes may include the entity-parallel relationship, the entity-fill relationship, and the entity-narrative relationship.
(2) **Lack of Representativeness in Training Data**: The Universum class is the complementary set of predefined target classes in the task. Therefore, it contains all possible implicit classes, i.e., classes not explicitly defined in the task but may appear in the real world. In this case, Universum samples in the training data are unable to sufficiently represent all possible patterns of the genuine distribution of the Universum class. As depicted in Figure 1(b), gray samples represent Universum samples in the test set that are not represented by the training data.

Despite the substantial difference between the target classes and the Universum class, this issue has long been neglected by the NLP research community. The majority of works (Zhu & Li, 2022; Ye et al., 2022; Wan et al., 2022; Pouran Ben Veyseh et al., 2022; Tian et al., 2021; Fu et al., 2021; Li et al., 2021b) treat the Universum class and target classes equally. Typically, a linear layer and a softmax function are applied at the end of the model to generate decision boundaries. The generated region of each class has open boundaries, which we believe are inappropriate for tasks containing the Universum class.

How to take into account the different properties of the Universum class and target classes in classifier design? In this work, we propose a closed boundary learning method for classification-based tasks with the Universum class. Traditional methods often adopt open boundary classifiers since the problem is under the closed-world assumption. However, the open decision boundaries can easily misclassify Universum samples, as illustrated in Figure 1(c). Therefore, we propose to use closed boundary classifiers as shown in Figure 1(d). We constrain the space of target classes to be closed spaces and designate the area outside all closed boundaries in the feature space as the space of the Universum class. This treatment perfectly fits the nature of the Universum class: a sample is marked

as the Universum if it does not belong to any target class during labeling. The aforementioned two properties are also well addressed in this way.

Although classifiers with closed boundaries are often used for problems under the open-world assumption, such as out-of-distribution (OOD) detection (Gomes et al., 2022; Sun et al., 2022), open set recognition (Zhang et al., 2021; Liu et al., 2020), and anomaly detection (Zong et al., 2018), our proposed method is different from these works in the following aspects. Firstly, we are addressing a different problem here: traditional classification under the closed-world assumption. The unique properties of the Universum class are overlooked by most studies in this field, and the classifier with closed decision boundaries is rarely used in traditional classification tasks containing the Universum class. Secondly, in a methodological point of view, our method differs from works in open-world classification in both problem setting and methodology, which are elaborated in the Related Work of this paper.

The main contributions of this work are summarized as follows:

- We bring attention to an important issue that is frequently neglected in classification-based NLP tasks like NER, RE, SA, and NLI: the Universum class, such as the *other* class, exhibits very different properties from target classes and should be treated differently.

- Methodologically, we generate closed boundaries with arbitrary shape, which include the commonly used spherical-shaped boundary as a special case. In addition, we leverage the information of both target classes and the Universum class to learn the decision boundary. We propose a boundary learning loss to generate the boundary based on the balance of misclassified Universum samples and the target class samples. In contrast to the intuitive two-step pipeline method, where Universum identification and multi-class classification may lead to error propagation, we develop a strategy to estimate the probability of the Universum class without relying on its intrinsic sample distribution and learn the classification of the Universum class and target classes jointly.

- We apply our closed boundary learning framework to 6 state-of-the-art (SOTA) works in 3 different classification-based tasks with Universum class. In all 6 experiments, the accuracy or F1-score of the model is improved over the original work.

2 Related Works

2.1 Classification Tasks with the Universum Class

The Universum class widely exists in classification based tasks in NLP, such as relation extraction (Zhang et al., 2017), named entity recognition (Tjong Kim Sang & De Meulder, 2003), and sentiment analysis (Jiang et al., 2019), as summarized in Table 1. It should be noted that the span-based methods (Zhu & Li, 2022; Li et al., 2021a) enumerate all possible spans and classify the entity from the whole set of spans, which introduces an extra *other* class. Despite the heterogeneity and lack of representativeness of the Universum class, current works (Zhu & Li, 2022; Wan et al., 2022; Fu et al., 2021; Tian et al., 2021; Chen et al., 2021; Li et al., 2021b; 2020; Yu et al., 2020) solve the classification problems containing the Universum class as normal multi-class classification problems and treat the Universum class and target classes equally.

2.2 Closed Boundary Learning Methods

Closed boundaries are often adopted in research fields of out-of-distribution (OOD) detection (Gomes et al., 2022; Ren et al., 2021; Chen et al., 2020), open set recognition (Zhang et al., 2021; Liu et al., 2020), anomaly detection (Zong et al., 2018), and outlier detection (Sharan et al., 2018; Sugiyama & Borgwardt, 2013). We adopt the terminology of generalized OOD detection (Yang et al., 2021b) to encompass these problems and analyze the differences between our work and generalized OOD detection.

2.2.1 DIFFERENCE IN PROBLEM SETTING

The problem we raised in this paper lies in **the intersection of open-world and closed-world assumptions.** Classification tasks can be categorized into problems based on closed-world assumption and open-world assumption.

- *Closed-world assumption*: The test data is assumed to be drawn from the same distribution as the training data, known as in-distribution (ID).
- *Open-world assumption*: In a real-world environment, out-of-distribution (OOD) samples that are not covered in the training data are expected in the test data.

The classification problem with the Universum class shares similarities and differences with both assumptions. The tasks highlighted in this paper, such as NER, RE, and SA, are problems treated under the closed-world assumption. Nevertheless, the Universum samples in these tasks are analogous to OOD samples in open-world assumption in two ways: multiple implicit classes are contained in the Universum/OOD class and there are unseen Universum/OOD samples in the test set. On the other hand, despite the similarity between the Universum and OOD samples, the problem we raised also differs significantly from open-world problems. This is because the OOD samples are not available in the training data in generalized OOD detection problems, whereas a considerable number of Universum samples are included in the training data in our problem setting. The information of existing Universum samples is important to generate accurate decision boundaries in our problem.

2.2.2 DIFFERENCE IN METHODOLOGY

By definition, the OOD detection problem assumes that the training data do not contain any OOD samples. However, a branch of the OOD studies, known as outlier exposure (Katz-Samuels et al., 2022; Ming et al., 2022; Yang et al., 2021a; Thulasidasan et al., 2021; Mohseni et al., 2020; Hendrycks et al., 2018), introduces auxiliary outlier data during training. The introduced auxiliary data makes it close to the format of our raised classification problems with the Universum class. However, outlier exposure methods are not suitable for our problem. The outlier exposure method mostly adopts a two-step approach that consists of multi-class classification and OOD identification. The OOD identification step is treated as a binary classification problem that distinguishes OOD and ID samples based on a score obtained by cross entropy (Yang et al., 2021a; Thulasidasan et al., 2021; Mohseni et al., 2020; Hendrycks et al., 2018), and energy function (Katz-Samuels et al., 2022; Ming et al., 2022). However, both cross entropy and energy function are monotonically varying. As a result, the decision boundary derived from a threshold score of the monotonically varying function is an open boundary, which leaves the heterogeneity and representativeness issues we pointed out in this paper still unresolved. In addition, the two-step approach will also suffer from error propagation.

From a methodological point of view, our work is also different from the works in generalized OOD using closed boundaries. In generalized OOD studies, the closed boundaries are formulated by the classic density-based method (Pidhorskyi et al., 2018; Hu et al., 2018), one-class classification method (Reiss et al., 2021; Ruff et al., 2018), or distance-based method (Gomes et al., 2022; Sun et al., 2022; Zhang et al., 2021; Zaeemzadeh et al., 2021; Shu et al., 2020). The distance-based methods are limited to spherical boundary shapes but our method can generate arbitrary shape boundaries. The one-class classification method formulates only one closed boundary between positive and negative samples while our work generates closed boundaries for every target class. The closed boundaries are formulated in the density-based method in cases where non-monotonic density functions are separated into the inner and outer regions by the threshold value. However, only positive samples are used to learn decision boundaries in density-based method, while both target class samples and Universum samples are used in our work.

3 METHOD

Problem Definition: The goal of our proposed method is to learn closed decision boundaries for target classes and jointly classify the Universum samples and target samples. In order to make our proposed method compatible with most existing classification methods, the starting point of our method is the representations of the final layer of classification models $\mathrm{H} = \{\mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_{n-1}\} \in \mathbb{R}^{n \times l}$.

Table 1: The tasks and datasets that the Universum class exists.

Task	Dataset	Label Name	Proportion
Relation Extraction	SemEval 2010 Task 8 (Hendrickx et al., 2019)	Other	17.4%
Relation Extraction	TARCED (Zhang et al., 2017)	No relation	79.5%
Named Entity Recognition	CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003)	Miscellaneous	14.6%
Named Entity Recognition (span based method)	CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003)	Other	>90%
Aspect Category Sentiment Analysis	MAMS (Jiang et al., 2019)	Neutral	43.4%

3.1 Defining the Universum Class

Universum class exists in many tasks and datasets as we summarized in Table 1. Notably, the Universum class has various names such as *other* and *miscellaneous*, etc. In sentiment analysis, the *neutral* class can be considered as the Universum class because the word *neutral* is defined as "having no strongly marked or positive characteristics or features", which means the *neutral* class is a collection of all samples without strong emotions. Similarly, the *no relation* class in the relation extraction task can be considered as the Universum class.

3.2 Dimension of the Final Layer

In our work, the output of the final layer of the model is served as representations instead of logits, and our proposed closed boundary classifier learns decision boundaries in the feature space of the final layer. Therefore, the last layer's dimensionality can affect the performance of the model. Recalling the classic Hughes phenomenon (Hughes, 1968) that the model accuracy is monotonically increasing first and then monotonically decreasing with the dimension of data, the dimension of the final layer can be carefully chosen to boost model performance.

3.3 Pretraining

We adopt N-pair loss (Sohn, 2016) for pretraining, with the goal of achieving small intra-class distance and large inter-class distance. Notably, in accordance with the nature of the Universum class, we make a change that does not require the model to reduce the intra-class distance of Universum samples during the pretraining.

3.4 Generating Closed Boundary of Arbitrary Shape for target classes

Existing closed boundary classification methods mainly use spherical shape boundaries (Zhang et al., 2021; Liu et al., 2020); however, we argue that the spherical shape may not be the optimal solution because data samples are unlikely to perfectly fit into a sphere, and a spherical shape boundary may produce misclassifications. We adopt the Gaussian mixture model (GMM) and the threshold value to generate boundaries with arbitrary shapes.

3.4.1 Gaussian Mixture Model

We apply GMM with m components to estimate the class conditional probability distribution for each target class C_i, and further derive the joint probability estimation for each class.

$$p(\mathbf{h}_k \mid C_i) = \sum_{i=1}^{m} \pi_i \mathcal{N}\left(\mathbf{h}_k; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right) \tag{1}$$

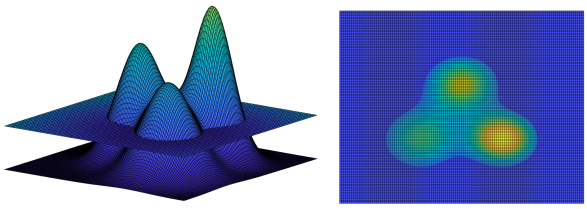
$$p(\mathbf{h}_k, C_i) = p(\mathbf{h}_k \mid C_i) p(C_i) \tag{2}$$

where $\mathbf{h}_k$ denotes the input feature vector of the kth sample, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the estimated mean vector and covariance matrix of the ith Gaussian components, respectively. π_i is the non-negative mixture weight under the constraint that $\sum_{i=1}^{m} \pi_i = 1$. $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}_i$, and π_i are all learnable parameters in the model.

According to Bayes Theorem, the posterior probability $p(C_i \mid \mathbf{h}_k) = \frac{p(\mathbf{h}_k|C_i)p(C_i)}{p(\mathbf{h}_k)}$. Since we are interested in $\operatorname{argmax}_{C_i} \frac{p(\mathbf{h}_k|C_i)p(C_i)}{p(\mathbf{h}_k)}$, the decision can be made based on joint probability $p(\mathbf{h}_k, C_i)$.

3.4.2 ARBITRARY SHAPE BOUNDARY

Geometrical View: Inspired by the DENCLUE algorithm in generating arbitrary shape clusters (Hinneburg & Keim, 1998), we introduce a threshold value ξ_i for each target class. A closed boundary of arbitrary shape is formulated by points satisfying $p(\mathbf{h}, C_i) = \xi_i$. Figure 2 is an illustration of formulating an arbitrary shape boundary in a three-dimensional space. A sample is assigned to class C_i if it is located inside the closed boundary. If the number of components of the GMM is set to one, then the shape of the boundary becomes spherical, with its center and covariance matrix specified by $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$, respectively. In this sense, the commonly used spherical shape boundary (Zhang et al., 2021; Liu et al., 2020) is a special case of our method.



(a) (b)

Figure 2: Illustration of generating arbitrary shape boundaries.

Probabilistic View: The above geometrical process can be described as:

$$\begin{cases} \mathbf{h}_k \in C_i \text{ if } p(\mathbf{h}_k, C_i) > \xi_i \\ \mathbf{h}_k \notin C_i \text{ if } p(\mathbf{h}_k, C_i) \leq \xi_i \end{cases} \tag{3}$$

3.5 PROBABILITY ESTIMATION FOR THE UNIVERSUM CLASS

3.5.1 MOTIVATION AND THE ESTIMATION

We classify the samples of Universum class and target classes based on the following rules in this work:

- *Rule 1*: A sample is assigned to the Universum class if it is not located inside any of the closed boundaries of target classes.
- *Rule 2*: A sample is assigned to the target class with the highest $p(\mathbf{h_k}, C_i)$ if it is located inside at least one closed boundary.

An intuitive way to incorporate the above rules is a two-step method: first, check whether the sample belongs to the Universum class according to Rule 1. Second, if the sample does not belong to the Universum class, assign the sample to the target class with the highest probability. However, such a pipeline method has the issue of error propagation. Another way is to use the distance relationship between the sample point and centers of target classes. However, the distance, e.g., Mahalanobis distance, is only a simple special case obtained from Section 3.4. In addition, the distance-based method is incompatible with cross-entropy loss, which has the benefit of being directly related to accuracy. Therefore, estimating the probability of the Universum class can allow us to jointly classify the Universum class and target classes, and is also compatible with cross-entropy loss.

However, general estimation methods, such as the GMM model, may not be a practical solution to estimating the probability distribution for the Universum class due to the lack of representativeness of training samples of the Universum class. A strategy that both complies with Rule 1 and facilitates the learning of the neural network needs to be devised.
For compliance with Rule 1, the estimated probability of the Universum class must be greater than the probability of all target classes and less than the probability of at least one target class for samples of the Universum class and target classes, respectively. Taking Equation 3 into account, a delicate function involving ξ_i is required to address this issue. To enhance the learning of neural networks, the gradient learned from an Universum sample should help move the sample away from its closest target class boundary. Hence, $\max(p(\mathbf{h}_k, C_i))$ is likewise needed for the estimation of probability distribution. Therefore, we propose to estimate the probability distribution of the Universum class as follows:

$$p(\mathbf{h}_k, U) = \lambda \frac{\xi_u^2}{p(\mathbf{h}_k, C_u)} + (1-\lambda) \frac{\xi_v^2}{p(\mathbf{h}_k, C_v)} \tag{4}$$

$$\lambda = \begin{cases} 1,\ p(\mathbf{h}_k, C_u) > \xi_u \\ 0,\ p(\mathbf{h}_k, C_u) \leq \xi_u \end{cases}, \begin{cases} u = \text{argmax}_i \frac{p(\mathbf{h}_k, C_i)}{\xi_i} \\ v = \text{argmax}_i\, p(\mathbf{h}_k, C_i) \end{cases} \tag{5}$$

ξ_i is the threshold value of target class i, and $u, v \in \{1, 2, \ldots, n-1\}$.

3.5.2 Analysis of the Proposed Estimation

For estimated probability of the Universum class in Equation 5, two cases are possible.
Case 1: $p(\mathbf{h}_k, C_u) > \xi_u$, i.e., sample $\mathbf{h}_k$ is located inside at least one closed boundary.
In this case, we have

$$p(\mathbf{h}_k, U) = \xi_u \frac{\xi_u}{p(\mathbf{h}_k, C_u)} < \xi_u < p(\mathbf{h}_k, C_u)$$

Since $p(\mathbf{h}_k, U) < p(\mathbf{h}_k, C_u)$, the model will select the target class i with the highest $p(\mathbf{h}_k, C_i)$, which fits perfectly with Rule 2.
Case 2: $p(\mathbf{h}_k, C_u) \leq \xi_u$, i.e., the sample $\mathbf{h}_k$ distribute outside every closed boundary.
Combining the condition of case 2 and Equation 5, we have

$$\forall i \in \{1, 2, \ldots, n-1\} : \frac{p(\mathbf{h}_k, C_i)}{\xi_i} \leq \frac{p(\mathbf{h}_k, C_u)}{\xi_u} \leq 1$$
$$\text{i.e.,} \forall i \in \{1, 2, \ldots, n-1\} : p(\mathbf{h}_k, C_i) \leq \xi_i \tag{6}$$

Combining Equation 5 and Equation 6, we can derive that:

$$\forall i \in \{1, 2, \ldots, n-1\} :\ p(\mathbf{h}_k, U) = \xi_v \frac{\xi_v}{p(\mathbf{h}_k, C_v)} \geq \xi_v \geq p(\mathbf{h}_k, C_v) \geq p(\mathbf{h}_k, C_i) \tag{7}$$

In case 2, from Equation 6 and Equation 3, we can learn that sample $\mathbf{h}_k$ is located outside all closed boundaries of target classes. In this case, the probability of Universum class $p(\mathbf{h}_k, U)$ obtains the largest value, as shown in Equation 7. Therefore, Rule 1 is perfectly expressed by the proposed probability estimation of the Universum class.

3.6 Boundary Learning Loss

To facilitate the learning of the closed decision boundaries, we propose a boundary learning loss below. The decision boundary should be adjusted to the balance of misclassified samples inside and outside the boundary. For example, if samples of class j distribute inside the boundary of class i, then the boundary should contract to exclude such samples and vice versa.

$$\mathrm{L_{bl}} = \frac{1}{M} \sum_{i=1}^{N-1} \left(\sum_{k \in \mathbb{O}} w_k \log \frac{\xi_i}{p(\mathbf{h}_k, C_i)} + \sum_{l \in \mathbb{I}} w_l \log \frac{p(\mathbf{h}_l, C_i)}{\xi_i} \right)$$

M is the total number of misclassified samples for all boundaries, N is the number of classes, $\mathbb{O}$ and $\mathbb{I}$ denote the set of samples misclassified outside and inside the decision boundary i, respectively. The weights in the loss function are $w_k = \frac{p(\mathbf{h}_k, C_i)}{p(\mathbf{h}_k, C_i) + \xi_i}$, $w_l = \frac{\xi_i}{p(\mathbf{h}_l, C_i) + \xi_i}$, and they are detached and cut off the gradient. Weights w_k and w_l have smaller values for samples located far from the boundary, enabling the boundary to be adjusted primarily on the basis of easily and semi-hard negatives instead of hard negatives.

4 Experiments

4.1 Datasets

Our proposed method is evaluated on six different SOTA models on three classic datasets in NLP, including SemEval 2010 Task 8 (Hendrickx et al., 2019) for relation extraction, MAMS (Jiang et al., 2019) for aspect category sentiment analysis, and CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003) for named entity recognition. The proportion of Universum samples in the three datasets are 17.4% (highest in 19 classes), 90%, and 43.4%. respectively. It is noteworthy that the ratio of

Table 2: The overall performance of applying closed boundary learning on baseline models.

Task	Method	F1/accuracy
NER	SpanNER (Fu et al., 2021)	92.09
	SpanNER (Fu et al., 2021) + COOLU	**93.50**
	BS (Zhu & Li, 2022)	92.53
	BS (Zhu & Li, 2022) + COOLU	**93.17**
RE	A-GCN (Tian et al., 2021)	88.67
	A-GCN (Tian et al., 2021) + COOLU	**89.33**
	TaMM (Chen et al., 2021)	88.76
	TaMM (Chen et al., 2021) + COOLU	**89.47**
ACSA	AC-MIMLLN (Li et al., 2020)	76.13%
	AC-MIMLLNN (Li et al., 2020) + COOLU	**77.35%**
	SCAPT (Li et al., 2021b)	84.13%
	SCAPT (Li et al., 2021b) + COOLU	**85.06%**

Universum class in the NER task is not calculated from the *miscellaneous* samples in the dataset but from the *other* samples which are introduced by the span-based method (Zhu & Li, 2022; Fu et al., 2021). Particularly, the span-based methods enumerate all possible spans for prediction, which lead to a large number of *other* samples.

4.2 Baselines and Evaluation Metrics

We evaluate the effectiveness of our proposed **Cl**osed b**O**undary **L**earning for classiciation with the **U**niversum class (**COOLU**) on 6 state-of-the-art works, including SpanNER (Fu et al., 2021), BS (Zhu & Li, 2022), A-GCN (Tian et al., 2021), TaMM (Chen et al., 2021), AC-MIMLLN (Li et al., 2020), and SCAPT (Li et al., 2021b). The implementation details are in the Appendix.
We employ the evaluation metrics used in the original models for comparison. Specifically, the accuracy is used in the task of ACSA, and the Macro F1 is used in tasks of NER and RE.

4.3 Overall Experimental Results

Table 2 shows the overall results for all 6 models. We reproduce the baseline models based on the officially released source code, and apply closed boundary learning on the source code. All reported results are the average of three runs. It should be noted that some results of baseline models are slightly different from those given in the original papers due to the variations in random seeds and package versions when reproducing baseline models from their officially released codes. Nevertheless, baseline models and models with closed boundary learning are fairly compared in our work under the same random seed and deep learning environment.
Models with our proposed closed boundary learning outperform the original models with open classifiers on NLP tasks containing the Universum class. The overall accuracy or F1 score is improved on all six models we evaluated, with the largest improvement from 92.09 to 93.50 in F1 score by simply adopting the closed boundary learning without changing the model.

4.4 A Closer Look at the Improvements

The improvement of overall performance is not only attributed to the improvement of the Universe class, but also to the improvement of all classes as a whole. We show the micro F1 score of each class by applying closed boundary learning on SpanNER in Table 3. The micro F1 score of *other* class, which is introduced by the span-based method, is not listed in the table because the overall F1 score of the task depends on the F1 score of the four entity types below. The micro F1 score is improved in all classes, with the absolute improvement of 2.43, 0.29, 1.31, and 4.02, respectively.
In addition, the precision score of each entity type is also improved significantly by adopting the closed boundary learning, with the largest absolute gain being 9.03. The relatively poor performance of the baseline method in precision score is consistent with our argument that the Universum class is easily misclassified if its special properties are neglected. The improvement in precision score indicates that our method can effectively avoid misclassification of the Universum class into target classes.

Table 3: The micro F1 score of SpanNER (Fu et al., 2021) with and without closed boundary learning.

Method	ORG			PER			LOC			MISC		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
SpanNER	89.92	89.81	89.87	97.74	96.47	97.11	93.05	93.88	93.46	78.99	82.83	80.87
SpanNER + COOLU	**94.29**	90.30	**92.30**	**98.54**	96.29	**97.40**	**96.17**	93.41	**94.77**	**88.02**	81.97	**84.89**

Table 4: The effect of pretraining on SpanNER (Fu et al., 2021) and SCAPT (Li et al., 2021b).

Method	F1/accuracy
SpanNER	92.09
SpanNER + N-pair pretraining	92.05
SpanNER + COOLU (N-pair pretraining)	**93.50**
SCAPT	84.13%
SCAPT + N-pair pretraining	84.16%
SCAPT + COOLU (N-pair pretraining)	**85.06%**

4.5 Ablation Study on N-pair Loss Pretraining

To further demonstrate the effectiveness of our closed boundary learning method and to rule out the possibility that the improvement of our model is due to the N-pair loss pretraining we adopted, we add an additional pretraining step to baseline models of SpanNER (Fu et al., 2021) and SCAPT (Li et al., 2021b). Table 4 indicates that the pretraining process alone cannot improve the accuracy of the original baseline models. The improvement brought about by our proposed closed boundary learning is not a result of the pretraining step but the result of the entire system.

4.6 The Impact of the Final Layer Dimension

We investigate the effect of last layer dimension on the accuracy of the model on the SpanNER (Fu et al., 2021) with closed boundary learning and present the result in Figure 3. The F1 score of the test set grows with increasing of dimensions and reaches a maximum value of 93.88 when the dimension is seven, and then decreases with the dimension. The trend fits well with the famous Hughes phenomenon Hughes (1968). Our method is not significantly affected by dimension and the overall result of SpanNER + COOLU reported in Table 2 is set as ten rather than the optimal value.

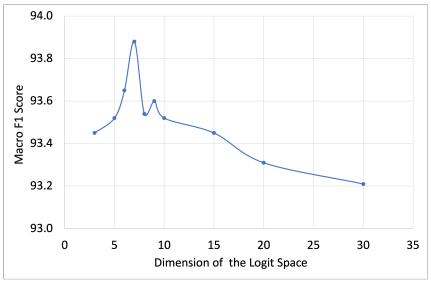


Figure 3: Illustration of generating arbitrary shape boundaries.

5 Discussion and Conclusion

In this work, we highlight an overlooked issue in classification-based NLP tasks that the Universum class is treated equally with target classes despite their significant differences. We propose a closed boundary learning method COOLU as a solution to this problem. Specifically, we generate closed boundaries with arbitrary shapes, develop a strategy to estimate the probability of the Universum class without relying on its intrinsic sample distribution, and propose a boundary learning loss to adjust decision boundaries based on the balance of misclassified samples inside and outside the boundary. Experimental results demonstrate that our proposed COOLU method improves the performance of all SOTA baseline models. The classification with the Universum problem that we emphasized brides the two distinct research areas of traditional classification and generalized OOD detection by exposing that the Universum class introduces open-world information to the task under the closed-world assumption. In addition, the closed decision boundaries can help the model to distinguish the knowledge boundary learned from the training data, which may contribute to the transferability of the model.

REFERENCES

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

Guimin Chen, Yuanhe Tian, Yan Song, and Xiang Wan. Relation extraction with type-aware map memories of word dependencies. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2501–2512, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.221. URL https://aclanthology.org/2021.findings-acl.221.

Xingyu Chen, Xuguang Lan, Fuchun Sun, and Nanning Zheng. A boundary based out-of-distribution classifier for generalized zero-shot learning. In *European Conference on Computer Vision*, pp. 572–588. Springer, 2020.

Jinlan Fu, Xuanjing Huang, and Pengfei Liu. SpanNER: Named entity re-/recognition as span prediction. 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)*, pp. 7183–7195, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.558. URL https://aclanthology.org/2021.acl-long.558.

Eduardo Dadalto Camara Gomes, Florence Alberge, Pierre Duhamel, and Pablo Piantanida. Igeood: An information geometry approach to out-of-distribution detection. In *International Conference on Learning Representations*, 2022.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid O Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. *arXiv preprint arXiv:1911.10422*, 2019.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018.

Alexander Hinneburg and Daniel A Keim. An efficient approach to clustering in large multimedia databases with noise. In *Knowledge Discovery and Datamining (KDD'98)*, pp. 58–65, 1998.

Weiming Hu, Jun Gao, Bing Li, Ou Wu, Junping Du, and Stephen Maybank. Anomaly detection using local kernel density estimation and context-based regression. *IEEE Transactions on Knowledge and Data Engineering*, 32(2):218–233, 2018.

Gordon Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory*, 14(1):55–63, 1968.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. A challenge dataset and effective models for aspect-based sentiment analysis. 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)*, pp. 6280–6285, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1654. URL https://aclanthology.org/D19-1654.

Julian Katz-Samuels, Julia B Nakhleh, Robert Nowak, and Yixuan Li. Training ood detectors in their natural habitats. In *International Conference on Machine Learning*, pp. 10848–10865. PMLR, 2022.

Fei Li, ZhiChao Lin, Meishan Zhang, and Donghong Ji. A span-based model for joint overlapped and discontinuous named entity recognition. 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)*, pp. 4814–4828, 2021a.

Yuncong Li, Cunxiang Yin, Sheng-hua Zhong, and Xu Pan. Multi-instance multi-label learning networks for aspect-category sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3550–3560, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.287. URL https://aclanthology.org/2020.emnlp-main.287.

Zhengyan Li, Yicheng Zou, Chong Zhang, Qi Zhang, and Zhongyu Wei. Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 246–256, Online and Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.22. URL https://aclanthology.org/2021.emnlp-main.22.

Bo Liu, Hao Kang, Haoxiang Li, Gang Hua, and Nuno Vasconcelos. Few-shot open-set recognition using meta-learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8798–8807, 2020.

Yifei Ming, Ying Fan, and Yixuan Li. Poem: Out-of-distribution detection with posterior sampling. In *International Conference on Machine Learning*, pp. 15650–15665. PMLR, 2022.

Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 5216–5223, 2020.

Stanislav Pidhorskyi, Ranya Almohsen, and Gianfranco Doretto. Generative probabilistic novelty detection with adversarial autoencoders. *Advances in neural information processing systems*, 31, 2018.

Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, Bonan Min, and Thien Nguyen. Document-level event argument extraction via optimal transport. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1648–1658, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.130. URL https://aclanthology.org/2022.findings-acl.130.

Tal Reiss, Niv Cohen, Liron Bergman, and Yedid Hoshen. Panda: Adapting pretrained features for anomaly detection and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2806–2814, 2021.

Jie Ren, Stanislav Fort, Jeremiah Liu, Abhijit Guha Roy, Shreyas Padhy, and Balaji Lakshminarayanan. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*, 2021.

Lukas Ruff, Robert Vandermeulen, Nico Goernitz, Lucas Deecke, Shoaib Ahmed Siddiqui, Alexander Binder, Emmanuel Müller, and Marius Kloft. Deep one-class classification. In *International conference on machine learning*, pp. 4393–4402. PMLR, 2018.

Vatsal Sharan, Parikshit Gopalan, and Udi Wieder. Efficient anomaly detection via matrix sketching. *Advances in neural information processing systems*, 31, 2018.

Yu Shu, Yemin Shi, Yaowei Wang, Tiejun Huang, and Yonghong Tian. P-odn: Prototype-based open deep network for open set recognition. *Scientific reports*, 10(1):1–13, 2020.

Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

Mahito Sugiyama and Karsten Borgwardt. Rapid distance-based outlier detection via sampling. *Advances in neural information processing systems*, 26, 2013.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *arXiv preprint arXiv:2204.06507*, 2022.

Sunil Thulasidasan, Sushil Thapa, Sayera Dhaubhadel, Gopinath Chennupati, Tanmoy Bhattacharya, and Jeff Bilmes. An effective baseline for robustness to distributional shift. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 278–285. IEEE, 2021.

Yuanhe Tian, Guimin Chen, Yan Song, and Xiang Wan. Dependency-driven relation extraction with attentive graph convolutional networks. 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)*, pp. 4458–4471, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.344. URL https://aclanthology.org/2021.acl-long.344.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003. URL https://aclanthology.org/W03-0419.

Juncheng Wan, Dongyu Ru, Weinan Zhang, and Yong Yu. Nested named entity recognition with span-level graphs. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 892–903, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.63. URL https://aclanthology.org/2022.acl-long.63.

Jason Weston, Ronan Collobert, Fabian Sinz, Léon Bottou, and Vladimir Vapnik. Inference with the universum. In *Proceedings of the 23rd international conference on Machine learning*, pp. 1009–1016, 2006.

Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8301–8309, 2021a.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021b.

Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4904–4917, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.337. URL https://aclanthology.org/2022.acl-long.337.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6470–6476, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.577. URL https://aclanthology.org/2020.acl-main.577.

Alireza Zaeemzadeh, Niccolo Bisagno, Zeno Sambugaro, Nicola Conci, Nazanin Rahnavard, and Mubarak Shah. Out-of-distribution detection using union of 1-dimensional subspaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9452–9461, 2021.

Hanlei Zhang, Hua Xu, and Ting-En Lin. Deep open intent classification with adaptive decision boundary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14374–14382, 2021.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pp. 35–45, 2017. URL https://nlp.stanford.edu/pubs/zhang2017tacred.pdf.

Enwei Zhu and Jinpeng Li. Boundary smoothing for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7096–7108, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.490. URL https://aclanthology.org/2022.acl-long.490.

Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018.

Table 5: The pretrained models chosen for each baseline model and the corresponding F1 score/accuracy reported in the original paper.

Method	Pretrained Model	Reported F1/accuracy
SpanNER (Fu et al., 2021)	BERT-base	92.28
BS (Zhu & Li, 2022)	RoBERTa	93.65
A-GCN (Tian et al., 2021)	BERT-base	89.16
TaMM (Chen et al., 2021)	BERT-base	89.18
AC-MIMLLN (Li et al., 2020)	Glove	76.42
SCAPT (Li et al., 2021b)	BERT-base	85.24

A APPENDIX

A.1 IMPLEMENTATION DETAILS

In the six baseline models we selected, different results based on multiple language models are often reported in one work. We choose one of the pretrained models used in each work, and reproduce the baseline models. The pretrained language model we used in each baseline and their reported results are summarized in Table 5.