FUSE: Measure-Theoretic Compact Fuzzy Set Representation for Taxonomy Expansion

Fred Xu, **Song Jiang**, **Zijie Huang**, **Xiao Luo**, **Shichang Zhang**, **Yuanzhou Chen**, and **Yizhou Sun**

Department of Computer Science, University of California, Los Angeles

fredxu@cs.ucla.edu

Abstract

Taxonomy Expansion, which models complex concepts and their relations, can be formulated as a set representation learning task. The generalization of set, fuzzy set, incorporates uncertainty and measures the information within a semantic concept, making it suitable for concept modeling. Existing works usually model sets as vectors or geometric objects such as boxes, which are not closed under set operations. In this work, we propose a sound and efficient formulation of set representation learning based on its volume approximation as a fuzzy set. The resulting embedding framework, *Fuzzy Set Embedding* (FUSE), satisfies all set operations and compactly approximates the underlying fuzzy set, hence preserving information while being efficient to learn, relying on minimum neural architecture. We empirically demonstrate the power of FUSE on the task of taxonomy expansion, where FUSE achieves remarkable improvements up to 23% compared with existing baselines. Our work marks the first attempt to understand and efficiently compute the embeddings of fuzzy sets.

1 Introduction

Taxonomy is a crucial data structure for modeling semantic concepts, hence of great importance for NLP (Lu et al., 2023; Xu et al., 2023; Yu et al., 2023). Concepts in a taxonomy can often be viewed as sets, the most fundamental object in mathematics, whose operations directly link to First Order Logic (FOL). For example, in a science taxonomy, "Biology" and "Computer Science" are semantic concepts, whose intersection results in a new concept "Bio-informatics", and "Diffusion Model" and "GAN" belong to a coarser-grained concept, "Generative Model". Usually, sets are seen as a *fixed collection* of objects. For example, the set $\mathbb{N}$ consists numbers $\{0, 1, \cdots\}$ by definition. However, in the context of semantic concepts, their meanings can change overtime and incorporate ambiguity. For example, "beauty" is a concept that has become broader overtime, and "deep learning models" can expand to have more elements with more discoveries made by the community. This underlying uncertainty and ambiguity are instead captured by a fuzzy set (Zadeh, 1999, 1978), an extension of classical sets.

A wide range of works have been developed for set representation learning. Early efforts are made to construct simple vector embeddings (Mikolov et al., 2013; Pennington et al., 2014; Devlin et al., 2019; Vaswani et al., 2023; Radford et al., 2018) based on similarity measures. To better model complex relationships such as asymmetrical relationships between concepts, geometric embeddings (Jiang et al., 2023; Hamilton et al., 2019; Ren et al., 2020; Ren and Leskovec, 2020) have been developed, which leverages the inherent geometric properties to model hierarchical relationships. However, these methods cannot address all the set operations including intersection, union, and complement. For example, box embedding (Jiang et al., 2023; Ren et al., 2020; Huang et al., 2023) doesn't define union and complement of boxes. Worse yet, existing geometric objects are not *closed* under set operations: the union of two boxes is no necessarily a box, which can compromise the consistency of reasoning in the embedding space.

In this paper, we directly tackle the challenge of fuzzy set representation learning for concept modeling. Our objective is to use their volume to quantify their information and their associated uncertainty. However, learning powerful representations for fuzzy sets is challenging. First, although extensive efforts have been made to incorporate deep learning techniques into fuzzy set modeling (Chen et al., 2022; Dasgupta et al., 2022b; Zhu et al., 2022), their training procedure could be expensive when the universe of discourse is large. Second, compared with geometric embeddings, which have

Findings of the Association for Computational Linguistics: ACL 2024, pages 2707–2720
August 11-16, 2024 ©2024 Association for Computational Linguistics

clear definitions of volume, it is unclear how to model the volumes of fuzzy sets due to the introduction of uncertainty and their abstract nature.

To tackle the previous challenges, we propose a principled and learnable model named *Fuzzy Set Embedding (FUSE)* for fuzzy set representation learning. The core of FUSE is to introduce a compact approximation of fuzzy sets and then prove that FUSE can arbitrarily approximate the original fuzzy sets under reasonable regularity conditions. FUSE avoids the computational burden of accounting for all the elements in the space of discourse at once, while enjoying the properties of fuzzy logic, hence satisfying all set operations. We further introduce a rank-based loss and asymmetric relations to enhance set representation learning. To validate the effectiveness of our proposed FUSE, we evaluate on taxonomy expansion task and show that FUSE can achieve the performance improvement up to 23% compared with state-of-the-art baselines, and we explore the effectiveness of our theoretical formulation through various ablation studies.

Our **main contributions** can be summarized as follows: (a) We propose an embedding framework to model fuzzy sets and show that the embeddings satisfy all set operations and are closed under set operations. (b) We systematically construct this embedding as a proper approximation of fuzzy sets. (c) We demonstrate the effectiveness of this embedding framework on taxonomy expansion by comparing it against previous vector and geometry-based embedding methods.

2 Related Work

2.1 Taxonomy Expansion

Taxonomy organizes concepts as a hierarchical graph, where nodes are concepts and edges denote "is-a" relationships between parent and child nodes. As new knowledge is emerging, taxonomy expansion seeks to expand existing taxonomy with new nodes, which is a fundamental task for many real-world applications such as information filtering and recommendation. Existing works have focused on using a lexical vector representation in the spirit of language modeling and word embedding (Chang et al., 2018; Snow et al., 2004; Mikolov et al., 2013; Pennington et al., 2014). More recently, geometric embeddings such as box embedding has been used to better model the asymmetric relationship between parent and child nodes (Jiang et al., 2023). Compared to vector-based representations, they improved both the predictive performance and interpretability of the learned embeddings.

2.2 Set Representation Learning

Set representation learning seeks to learn low-dimensional representations of data with a notion of volume and coverage. It is desirable when the representations can capture the rich semantic information and the complex relationships of data (Rossi et al., 2020; Wang et al., 2021; Zhang et al., 2022; Zhong et al., 2023). For example, language modeling (Devlin et al., 2019; Vaswani et al., 2023; Radford et al., 2018) has aimed to learn vectors to represent combinatorically intractable combinations of human languages. In this context, semantic concepts can be viewed as sets. Recently, geometry-based approaches (Ren et al., 2020; Dasgupta et al., 2022b; Ren and Leskovec, 2020; Chen et al., 2021) have further improved the efficiency of the representations by enabling set operation such as intersection, but they fail to cover all operations and are not closed under them. Fuzzy set theory has explicitly formulated a way to represent the ambiguity of sets such as concepts in taxonomy construction, while automatically satisfying all desired properties of sets (Chen et al., 2022; Zhu et al., 2022). It is an extension to classical set theory with extensive applications (van Krieken et al., 2022; Wagner and d'Avila Garcez, 2022; Liang et al., 2023; Yu et al., 2022; Xu et al., 2022). For example, Michael Boratko and McCallum (2022) and Dasgupta et al. (2022b) have explored the connection between fuzzy sets and box embeddings to model words. However, existing fuzzy set representations lack a principled approach on what the low-dimensional representation stands for, and can be inefficient when the number of sets to model increases. We propose a novel solution by identifying the central characterization of a fuzzy set as its volume and approximate it using a compact representation, while yielding superior performance on the set representation learning task of taxonomy expansion.

3 Preliminary

3.1 Fuzzy Sets

In contrast to classical set theory, which assigns a Boolean value to whether an element belongs to a set, a fuzzy set (Zadeh, 1978) assigns a value between 0 and 1 to denote a *degree of membership*. For a universe of discourse U, a fuzzy set is mathematically defined as a tuple $A = (U, m_A)$, where $A \subseteq U$ and $m_A : U \to [0, 1]$ is its membership function. For any element $x \in U$, $m_A(x)$ represents the degree of membership of element x in A. Fuzzy set models the uncertainty of membership by encoding imprecision and ambiguity in concepts. As an example, it can be used to describe the compatibility between two concepts, such as "is-a" relationship. For example, for a concept "Kobe Bryant" and a set of entities {Basketball Player, Team Owner, Entrepreneur}, a fuzzy membership function can be represented as the set $\{1.0, 0.1, 0.9\}$, each signifying "Kobe Bryant"'s compatibility with each of the concepts in the set.

Similar to standard sets, intersection, union, and complement between fuzzy sets are defined. Fuzzy set is related to fuzzy logic, which defines logical operations over soft truth values and follows Gödel, product, or Łukasiewicz systems. For a detailed discussion of fuzzy logic systems, see Chen et al. (2022). For language modeling, fuzzy sets can be used to model the ambiguity of the semantic meanings of words (Dasgupta et al., 2022a). In taxonomy, fuzzy sets can be used to model concepts.

3.2 Possibility Theory

The membership function m_A associated with a fuzzy set A is constructed based on the theory of possibility in Zadeh (1999, 1978). In the formulation of Zadeh (1978), to reason about linguistic concepts such as "likely", a fuzzy set can be endowed with a probability-possibility distribution:

Definition 1 (**Possibility-Probability Distribution**). *Let U be the universe of discourse, and $(U, \mathcal{F}, P)$ be a probability space, where $\mathcal{F}$ is the sigma-algebra and P is the probability measure. Let X be a fuzzy variable that can take any values $x \in U$, and let F be a fuzzy subset of U with membership function m_F, then the* ***possibility of probability*** *of X with respect to F is:*

$$\pi_{P,X} = \int_U \pi_X dP = \int_U m_F dP.$$

where π_X is the possibility distribution associated with X and $\int_U m_F dP$ is the Lebesgue integral of the membership function w.r.t to the probability measure P.

Here the Lebesgue integral $\int_U m_F dP$, in the sense of fuzzy set theory, represents the amount of information contained by the fuzzy set F. This construct can be seen as the measurement of information and uncertainty in the fuzzy variable X, making it a desirable quantity to approximate when learning a low-dimensional embedding of a fuzzy set. In our Fuzzy Set Embedding, we generalize definition 1 in definition 4.

4 Proposed Framework: FUSE

We now present **Fuzzy Set Embedding (FUSE)** for learning set representations in a principled way.

4.1 Fuzzy Set Embedding

To construct a proper embedding for fuzzy sets, we assume that the universe of discourse U admits a finite partition, $\{U_i\}_{i=1}^d$, such that $U = \bigcup_{i=1}^d U_i$, and U_i, U_j are disjoint if $i \neq j$. For a formal description of this assumption, see Appendix B. In particular, this indicates that fuzzy set membership function m_A has an associated *simple function*:

Definition 2 (**Simple Fuzzy Set**). *Let $(U, \mathcal{F}, \xi)$ be a measure space and Let $U = \bigcup_{i=1}^d U_i$ be a finite partition of the universe U, and let $A \in \mathcal{F}$ and m_A its membership function, then the* ***Simple Fuzzy Set*** *associated with A is the tuple (U, μ_A), where:*

$$\mu_A(x) := \sum_{i=1}^{d} \mathbf{1}_{\{x \in U_i\}} \mu_A^{(i)}(x), \forall x \in U \quad (1)$$

is the ***Simple Membership Function*** *of A, where $\mathbf{1}$ is the indicator function and $\forall x \in U, \forall i \in \{1, \cdots, d\}$:*

$$\mu_A^{(i)}(x) := \sup_{x \in U_i} m_A(x). \quad (2)$$

μ_A can be summarized in d values $\mu_A^{(1)}, \cdots, \mu_A^{(d)}$, each determined by the supremum of m_A in the corresponding partition. To facilitate the standard application in deep learning, we formulate an alternative representation in vector form to distinguish it from the functional representation denoted in Eqn. 1.

Definition 3 (**Fuzzy Set Embedding**). *Let $A = (U, \mu_A)$ be a simple fuzzy set defined on the measure space $(U, \mathcal{F}, \xi)$, where $U = \bigcup_{i=1}^d U_i$, then its corresponding* ***Fuzzy Set Embedding*** *(FUSE) is the d-dimensional vector:*

$$\mathcal{U}_A := [\mu_A^{(1)}, \cdots, \mu_A^{(d)}], \quad (3)$$

Since we are reducing the reasoning space from $[0, 1]^{|U|}$ to $[0, 1]^d$, we need to examine the loss incurred by this reduction. To reason about it in detail, we provide the following definition inspired

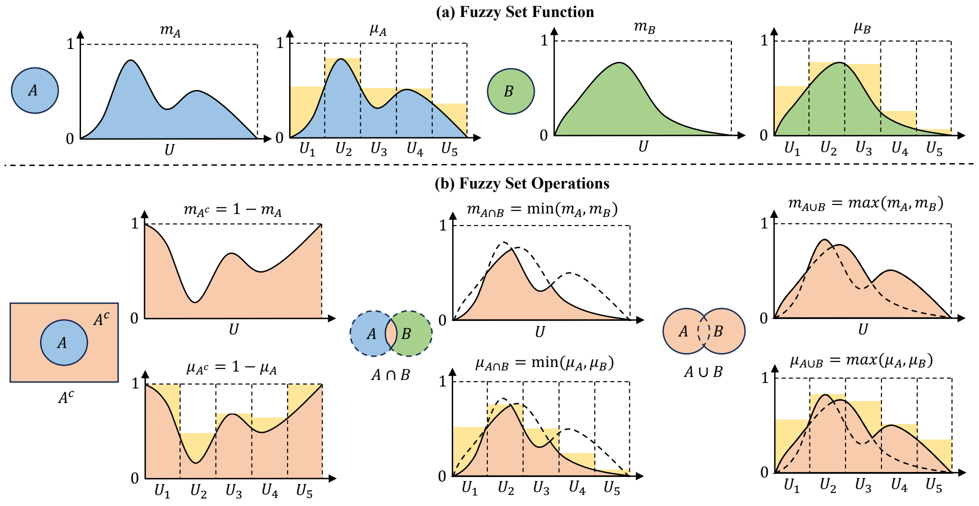

Figure 1: Illustration of set operations under fuzzy set membership function for two sets $A, B \in U$. m_A is the fuzzy set membership function and μ_A the corresponding simple membership function (Definition 2). By using the fuzzy set representation, all the set operations (intersection, union, complement) are well-defined (Gödel definition is used for easier illustration), and after set operations, results are still fuzzy set. For illustration, here we partition the universe U into 5 partitions $\{U_1, \cdots, U_5\}$.

by Zadeh (1978); Nahmias (1978) to quantify the amount of information covered by the fuzzy sets across the entire universe U.

Definition 4 (Simple Fuzzy Measure). *Let U be a compact universe of discourse and let $A = (U, m_A)$ be a fuzzy subset of U. Let $(U, \mathcal{F}, \xi)$ be a measure space defined on U, then the **fuzzy measure** of the fuzzy set (U, m_A) is:*

$$\mathbb{P}(A) := \int_U m_A d\xi.$$

*Then a **Simple Fuzzy Measure** of a simple fuzzy set $A = (U, \mu_A)$ is defined as:*

$$\mathbb{P}_\mu(A) := \int_U \mu_A d\xi.$$

Given a finite partition, furthermore:

$$\mathbb{P}_\mu(A) = \sum_{i=1}^{d} \int_{U_i} \mu_A^{(i)} d\xi = \sum_{i=1}^{d} \mu_A^{(i)} \xi(U_i), \quad (4)$$

where $\xi(U_i)$ corresponds to the measure of partition set U_i. If ξ is a probability measure, then Definition 4 corresponds to Definition 1. In practice, we examine choices of different measures empirically in Section 5. In short, a simple fuzzy set $A = (U, \mu_A)$ approximates the fuzzy measure of the underlying fuzzy set (U, m_A). We state this observation formally in theorem 1 and illustrates it in Figure 2(a).

Theorem 1. *Let U be a compact universe of discourse and $(U, \mathcal{F}, \xi)$ a measure space. Let A be a fuzzy subset of U and m_A its membership function that's measurable. Moreover, let μ_A be its simple membership function, then $\forall \epsilon > 0$, $\exists \delta > 0, d > 0$ such that if $d\delta = \xi(U)$ and $||U_i|| := \min_i \xi(U_i) < \delta$, we have:*

$$0 < \mathbb{P}_\mu(A) - \mathbb{P}(A) < \epsilon.$$

The fuzzy measure of a simple fuzzy set is an upperbound for the fuzzy measure of its underlying fuzzy set and converges to the possibility of its underlying fuzzy set when the partition is sufficiently fine-grained. With suitable assumption on the function m_A, we can further establish the rate of convergence:

Corollary 1. *Following the condition in definition 4, if in addition m_A is Lipschitz-continuous or of bounded variation, then the convergence rate in 1 is $O(1/n)$, where n is the number of partitions.*

4.2 Embedding-based Fuzzy Set Operators

FUSE combines set theory and measure theory and provides a theoretically sound embedding. For an entity $x \in U$, we treat it as a concept and associate with it a fuzzy subset (U, m_A), representing its compatibility with other concepts. We can treat every entity as a fuzzy set embedding, define set operations in the language of set theory, and compute them using vector operations. Suppose we have two entities $x, y \in U$, and $\mathcal{U}_A, \mathcal{U}_B \in [0, 1]^d$

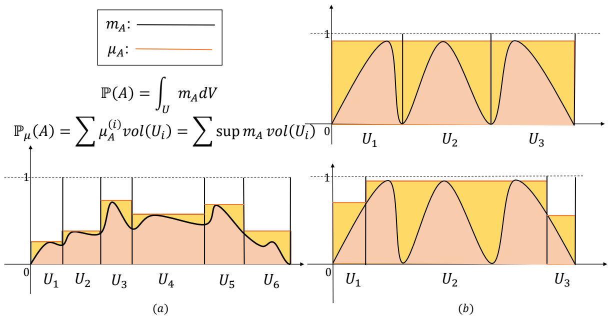

Figure 2: On the left plot (a), we illustrate when the universe is a compact subset of $\mathbb{R}$. Fuzzy measure $\mathbb{P}$ is the Riemann Integral of m_A on embedding space $\mathcal{U}$ (the orange region), while the Simple Fuzzy Measure $\mathbb{P}_\mu$ is its upper Darboux sum (the yellow region). On right plot (b), we demonstrate that for a fixed number of partitions, different choices of partition size result in different approximation of $\mathbb{P}$: the bottom-right partition has better approximation than the top-right partition, since it results in less over-estimation.

are the two fuzzy set embeddings associated with them, we can define following operations:

- **Fuzzy Mapping**: Every entity/element is a singleton set, and $\mathcal{M}$ maps an entity $x \in U$ to its associated fuzzy set embedding $\mathcal{U}_{\{x\}} \in [0,1]^d$. In the case of taxonomy expansion, the input is the word vector $\mathbf{x} \in \mathbb{R}^e$ obtained from a pre-trained language model like BERT (Devlin et al., 2019). To construct a map between the word vector and its associated fuzzy set embedding, we use a neural networks $f : \mathbb{R}^e \to [0,1]^d$:

$$\mathcal{U}_A = \mathcal{M}(A;\theta) = \sigma(f(\mathbf{x};\theta)) \in [0,1]^d, \quad (5)$$

  where σ is a normalization constraint to make the embedding space compact, such as sigmoid, 0-1 clamping, or Layernorm (Ba et al., 2016).

We can futher define set operations by product logic. Other fuzzy systems, such as Gödel logic, is illustrated in Fig. 1.

- **Intersection**: The intersection between the two fuzzy sets $A \cap B$ can be computed by element-wise product t-norm (Klement et al., 2013):

$$\mathcal{U}_{A \cap B} = \mathcal{U}_A \odot \mathcal{U}_B. \quad (6)$$

  where $\odot$ is element-wise multiplication.

- **Union**: The union of two fuzzy sets $A \cup B$ can be computed by element-wise product t-conorm:

$$\mathcal{U}_{A \cup B} = \mathcal{U}_A + \mathcal{U}_B - \mathcal{U}_A \odot \mathcal{U}_B. \quad (7)$$

- **Complement**: The complement of a fuzzy set A denoted as A^c can be computed as:

$$\mathcal{U}_{A^c} = \mathbf{1} - \mathcal{U}_A. \quad (8)$$

4.3 Taxonomy Expansion with FUSE

In this part, we use *taxonomy expansion* task to showcase the advantages of representing concepts with fuzzy set embeddings.

Membership Prediction with FUSE. After representing a concept with fuzzy sets, the core task of taxonomy expansion is to determine whether an element y belongs to a set A, and this is often used as a score function in pair-based relationship in taxonomy expansion task (Jiang et al., 2023; Shen et al., 2020; Yu et al., 2020). Using our framework, for some element $y \in U$, we can apply the entity mapping function $\mathcal{M}$ to find its fuzzy set embedding $\mathcal{U}_{\{y\}}$. Then we can simply measure the degree of membership of element y in some other fuzzy set A by considering the fuzzy measure of the fuzzy set embedding $\mathcal{U}_{A \cap \{y\}} = \mathcal{U}_A \odot \mathcal{U}_{\{y\}}$, which we denote as $\mathbb{P}_\mu(A \cap \{y\})$ and compute it as:

$$\mathbb{P}_\mu(A \cap \{y\}) = \sum_{i=1}^{d} \left(\mathcal{U}_A^{(i)} \mathcal{U}_{\{y\}}^{(i)} \right) \xi(U_i). \quad (9)$$

In training, we model the volume using global trainable weights $\mathbf{w} = \{w_1, \cdots, w_d\}$ with a normalization transform to restrict the type of measure ξ. Therefore, we approximate $\mathbb{P}_{A \cap \{y\}}$ and define the standard score function:

$$\psi(y, A) = \sum_{i=1}^{d} \mu_{A \cap \{y\}}^{(i)} w_i = (\mathcal{U}_A \odot \mathcal{U}_{\{y\}})^T \mathbf{w}. \quad (10)$$

and the corresponding ranking-based loss:

$$L(y, A) = -\log \sigma\left(\psi(y, A) - \gamma_p\right) - \frac{1}{k} \sum_{i=1}^{k} \log \sigma\left(\gamma_n - \psi(y', A)\right) \quad (11)$$

where (y, A) are positive pairs and (y', A) negative pairs, and γ_p, γ_n are margins for positive and negative predictions. We use different margins since by Theorem 1, fuzzy measure of the fuzzy set embedding is an upperbound for the underlying fuzzy set, so the result we obtain is overestimating the actual fuzzy measure. We provide an ablation study regarding choice of margin in section 5.

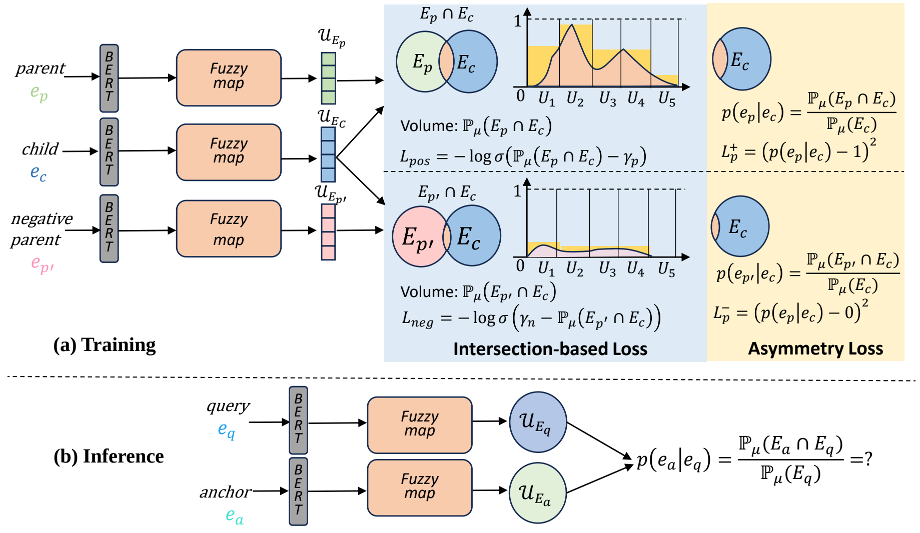
parent
e_p
child
e_c
negative parent
$e_{p'}$
BERT
Fuzzy map
$\mathcal{U}_{E_p}$
$\mathcal{U}_{E_c}$
$\mathcal{U}_{E_{p'}}$
$E_p \cap E_c$
E_p
E_c
1
0
U_1 U_2 U_3 U_4 U_5
Volume: $\mathbb{P}_\mu(E_p \cap E_c)$
$L_{pos} = -\log\sigma(\mathbb{P}_\mu(E_p \cap E_c) - \gamma_p)$
$E_{p'} \cap E_c$
$E_{p'}$
Volume: $\mathbb{P}_\mu(E_{p'} \cap E_c)$
$L_{neg} = -\log\sigma\left(\gamma_n - \mathbb{P}_\mu(E_{p'} \cap E_c)\right)$
$p(e_p|e_c) = \frac{\mathbb{P}_\mu(E_p \cap E_c)}{\mathbb{P}_\mu(E_c)}$
$L_p^+ = (p(e_p|e_c) - 1)^2$
$p(e_{p'}|e_c) = \frac{\mathbb{P}_\mu(E_{p'} \cap E_c)}{\mathbb{P}_\mu(E_c)}$
$L_p^- = (p(e_p|e_c) - 0)^2$
(a) Training
Intersection-based Loss
Asymmetry Loss
(b) Inference
query
e_q
anchor
e_a
$\mathcal{U}_{E_q}$
$\mathcal{U}_{E_a}$
$p(e_a|e_q) = \frac{\mathbb{P}_\mu(E_a \cap E_q)}{\mathbb{P}_\mu(E_q)} = ?$

Figure 3: The overview of FUSE used to model concepts in taxonomy. First the entities are converted into vectors using Bert, then the main Fuzzy Map transforms the vector into a fuzzy set embedding. (a) Training: FUSE is learned using both a volume-based intersection loss and a volume-based asymmetry loss. (b) Inference: use the taxonomy probability score to check the degree of containment of a query in the anchor. Here e are entities, E are their associated fuzzy sets, and $\mathcal{U}$ the fuzzy set embeddings.

Incorporating Asymmetric Relation. As signified in Jiang et al. (2023), membership prediction in a taxonomy expansion task usually involves asymmetric relations. For example, parent nodes in a taxonomy usually strictly incorporate the concept in the child nodes. Since the score function, which uses the intersection between two sets, is symmetric, here we propose another score function to signify this asymmetry. Suppose we have two entities $e_p, e_c \subseteq U$, such that e_c is a child of e_p. Let E_p, E_c be their associated fuzzy sets, then we can model this relationship by:

$$P(e_p|e_c) = \frac{\mathbb{P}_\mu(E_p \cap E_c)}{\mathbb{P}_\mu(E_c)} = \frac{(\mathcal{U}_{E_p} \odot \mathcal{U}_{E_c})^T \mathbf{w}}{\mathcal{U}_{E_c}}, \tag{12}$$

where we use the simple fuzzy measure for each set as its volume and use the ratio between the volume of $E_p \cap E_c$ and the volume of E_c as the result. For a positive child-parent pair (e_c, e_p), the loss is:

$$L_p^+ = (P(e_p|e_c) - 1)^2, \tag{13}$$

whereas for a negative child-parent pair (e_c, e_p'):

$$L_p^- = (P(e_p'|e_c) - 0)^2. \tag{14}$$

The main difference between Eqn. 10 and the case of box embedding in Jiang et al. (2023) is that the volume of a fuzzy set embedding spans the entire universe U. We combine the pair-based ranking loss and the asymmetric child-parent pair loss:

$$L_{taxo} = L(e_c, e_p) + \lambda(L_p^+ + L_p^-), \tag{15}$$

where λ is a hyper-parameter to control the strength of each loss.

Dataset	Environment			Science		
Metric	ACC	MRR	Wu&P	ACC	MRR	Wu&P
TAXI	16.7	N/A	44.7	13.0	N/A	32.9
HypeNet	16.7	23.7	55.8	15.4	22.6	50.7
BERT+MLP	11.1	21.5	47.9	11.5	15.7	43.6
TaxoExpan	11.1	32.3	54.8	27.8	44.8	57.6
STEAM	36.1	46.9	69.6	<u>36.5</u>	<u>48.3</u>	<u>68.2</u>
BoxTaxo	<u>38.1</u>	<u>47.1</u>	<u>75.4</u>	31.8	45.3	64.7
FUSE	**42.3**	**58.3**	**77.6**	**39.9**	**52.9**	**73.4**
FUSE (λ = 1.0)	**43.1**	**53.3**	74.3	**43.5**	**56.6**	**77.5**

Table 1: Results on taxonomy expansion compared to existing methods. Here bold font refers to the best performance results (compared to baseline) while underline refers to the second-best performance result. The results are reported as average over 5 runs. "N/A" is present since MRR is not applicable to TAXI. FUSE is our base model while FUSE (λ = 1.0) is the model with balanced weights for intersection and asymmetric loss.

5 Experiments

Dataset: We use two public datasets (Environment, Science) from SemEval-16 taxonomy construction tasks. Following the training setup in Jiang et al. (2023), we sample 20% of the leaf nodes as test set and use the rest as training data. The performance of taxonomy expansion task can be found in table 1, where the results are averaged over 5 runs to reduce variance.

Metrics: We use three metrics, Accuracy (ACC), Mean Reciprocal Rank (MRR), and Wu & Palmer similarity (Wu&P) to measure the performance for the baseline models and FUSE.

Baselines: The baseline comparison models are vector-based models like TAXI (Panchenko et al., 2016), HypeNet (Shwartz et al., 2016), BERT+MLP (Yu et al., 2020), TaxoExpan(Shen et al., 2020), STEAM (Yu et al., 2020), and geometry-based model like BoxTaxo (Jiang et al., 2023).

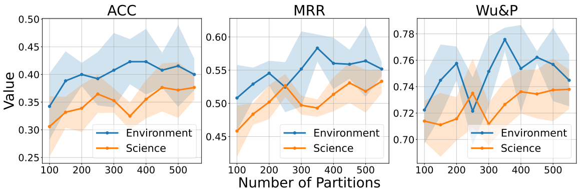

Figure 4: Trend of model performance with varying number of partitions on the science dataset.

5.1 Does FUSE Benefit from More Partitions?

Here we examine the impact of choice of number of partitions, since according to theorem 1, as d increases, we should expect better approximation to the underlying fuzzy set. In this experiment, we vary the number of partitions from 100 to 550, with increment of 50, and measure the performance of taxonomy expansion on both Science and Environment datasets, averaged over 5 runs. The resulting trend on both datasets can be seen in Figure 4.

In both cases, we can see a general upward trend on model's average performance when the number of partitions goes up. This provides empirical support for the theoretical result in theorem 1. The variance of model performance tends to increase when the number of partitions goes up, which suggests that smaller learning rates and other normalization for optimization may be considered.

Dataset	Environment			Science		
Metric	ACC	MRR	Wu&P	ACC	MRR	Wu&P
FUSE-sigmoid	**45.0**	57.2	75.6	**40.0**	52.7	**74.7**
FUSE-softmax	7.3	17.3	51.9	21.3	32.0	65.2
FUSE-01	41.3	55.8	75.4	37.3	52.3	72.3
FUSE	42.3	**58.3**	**77.6**	39.9	**52.9**	73.4

Table 2: Results on taxonomy expansion compared under different choices of normalization on volume weights. This corresponds to different choices of measure space.

5.2 Does the Choice of Measure Affect Model Performance?

As for the proposed new score function based on the volume of the fuzzy set in Eqn. 10, we study the impact of different choices of measure ξ. This corresponds to different choices of normalization applied to the volume of each partition. If we follow definition 1, then partition volumes follows a probability distribution, indicating a softmax normalization on the global weights. Otherwise, we can choose to use sigmoid or 0-1 clamping. Results over 5 runs for different choices of measures can be found in table 2. We observe that the softmax normalization, which enforces the volume weights to follow a probability distribution, doesn't work well overall. This suggest that the construction we proposed in definition 4 is more suitable than the classical definition in 1. We also observe that using a sigmoid normalization on the volume weights can improve results.

5.3 Does Asymmetry Loss Help Taxonomy Expansion Task?

In this ablation study, we examine the impact of applying asymmetry losses, since set intersection is a symmetric operation, whereas the membership relation is asymmetric. The value of λ in Eqn. 12 controls how much should the asymmetry-based loss affect the training (the greater the value of λ, the greater the impact). The results, averaged over 5 runs is in figure 5. We can see a trend of performance improvement on Science dataset and an increase in performance for $\lambda > 0$ for the Environment dataset. This result validates the importance of modeling asymmetric relations.

5.4 Does Wider Margin Affect Model Performance?

In this ablation study, we examine the impact of different margins on the learning performance, as presented in Eqn. 11, since theoretically, our construction of FUSE over-estimates the volume under the fuzzy set. In figure 6, which presents the average result over 5 runs. From the result, we can see that having wider margin (in this case $\gamma_p = 0.6, \gamma_n = 0.4$) does benefit the model performance, supporting our hypothesis that FUSE over-estimates the volume of a fuzzy set.

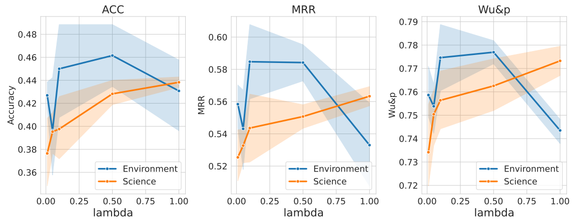

Figure 5: Model performance with different strength of asymmetry, lambda.

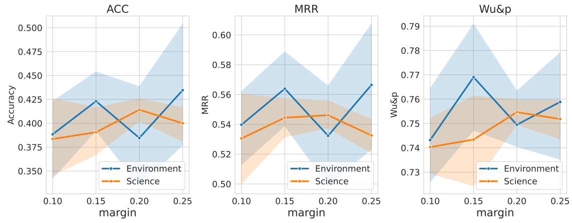

Figure 6: Model performance with different choice of margins.

5.5 Additional Experiment: Is there Synergy Between Different Modeling Choices?

In this part, we examine the synergy between configurations across all the ablation studies performed. In this case we study the case where all the best configurations ($\lambda = 0.5, \Delta\gamma = 0.2$, sigmoid-normalization) are used. We call the resulting model FUSE-comb(ine) and report the performance averaged over 5 runs compared to other models in table 3. The result doesn't immediate suggest that combining multiple best case scenarios would result in optimal performance. The strongest performance so far comes from setting $\lambda = 0.5$.

Table 3: Results on FUSE-comb compared with each individual best configurations

Dataset	Environment			Science		
Metric	ACC	MRR	Wu&P	ACC	MRR	Wu&P
FUSE-sigmoid	45.0	57.2	75.6	40.0	52.7	74.7
FUSE ($\lambda = 0.5$)	**46.2**	**58.4**	**77.7**	**42.8**	**55.1**	76.3
FUSE ($\Delta\gamma = 0.2$)	38.5	53.2	74.9	41.4	54.6	75.5
FUSE-comb	41.1	53.0	76.4	42.4	54.3	**76.7**

5.6 Infer about Union and Complement Using Trained Embeddings

In our taxonomy expansion experiment, the model is trained using only the volume-based intersection and asymmetry losses, without observing pairs of sets that are related by union or complements. In this case study we examine whether our embedding can generalize to these two operations.

Infer about Set Union: For a parent entity e_p in the taxonomy and its m child entities $e_{c_1}, \cdots, e_{c_m}$, we examine the similarity of union of fuzzy set embeddings of child entities with the fuzzy set embedding of the parent entity. That is, between $\mathcal{U}_{\bigcup_{i=1}^{m} E_{c_i}}$ and $\mathcal{U}_{E_p}$. To this end, we use the trained fuzzy set embedding from the FUSE ($\lambda = 1$) model and apply union operation in Eqn. 7 among all the child fuzzy set embeddings, then we rank the Euclidean distance between the obtained fuzzy set embedding (union of all child embeddings) against all the existing parent embeddings in the dataset. From 4, we observe that fuzzy set embedding captures union patterns. As an example from the Science dataset, for child entities ["calculus of variations", "analysis", "integral calculus"], with parent entity "calculus", the top-3 closest simple fuzzy set embedding corresponds to entity "calculus", "analysis", "geophysics", and the model's prediction is closest to the correct parent.

Infer about Set Complement: In this case we examine complement by the set operation $A \setminus B = A \cap B^c$. In particular, the fuzzy set for parent (denote it A) minus a child fuzzy set (denote it B) should be similar to the union of the remaining children fuzzy set. We follow the union, complement, and intersection operation to compute the fuzzy set embedding in this case, and again we use the embedding from FUSE ($\lambda = 1$) model. We compute the Euclidean distance between $A \cap B^c$ and all the existing child embeddings in the dataset. Here we present MRR and accuracy result in table 4. In contrast to union, it seems that complement doesn't achieve a reasonable performance. This may be due to the fact that the complement of a fuzzy set is taken over the entire universe of discourse, rather than simply in the scope of all the children entities.

Dataset	Environment		Science	
Metric	ACC	MRR	ACC	MRR
Union Inference	81.3	85.3	85.4	89.5
Complement Inference	2.9	14.8	11.9	26.6

Table 4: Results on Union and Complement Inference using FUSE trained only with intersection based loss

6 Conclusion

For taxonomy expansion, We propose a novel and theoretically sound Fuzzy Set Embedding (FUSE) to model concepts and relationship between concepts that incorporate set operations (intersection, union, complement). We show theoretically that FUSE preserves the information of the fuzzy set with sufficiently fine-grained partitions and demonstrate empirically that it can outperform existing vector-based and geometry-based embedding methods on taxonomy expansion. For future works, we believe that expanding the taxonomy dataset with more complicated combination of set operations, such as First Order Logic (FOL), can further improve the model performance.

7 Limitations

This work is the first attempt to use fuzzy set to model concepts in taxonomy expansion. To examine only the effectiveness of fuzzy set representation, we only use simple neural architectures and use only the child-parent pairs. The full capacity of FUSE should be further examined using datasets that contain First Order Logic (FOL) statements, since by construction, fuzzy sets should satisfy all the fuzzy logic axioms (Chen et al., 2022). This suggests future directions to expand taxonomy datasets with more complicated queries, and to handle more graph-structured data in social analysis and text mining (Ren and Leskovec, 2020; Ren et al., 2020; Chen et al., 2022; Zhu et al., 2022; Ju et al., 2023). Moreover, we can explore more explicit form of fuzzy membership function, such as a mixture of Gumbel boxes, to make the learning more concrete.

References

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization.

Haw-Shiuan Chang, ZiYun Wang, Luke Vilnis, and Andrew McCallum. 2018. Distributional inclusion vector embedding for unsupervised hypernymy detection.

Xuelu Chen, Michael Boratko, Muhao Chen, Shib Sankar Dasgupta, Xiang Lorraine Li, and Andrew McCallum. 2021. Probabilistic box embeddings for uncertain knowledge graph reasoning. In *Proceedings of the 19th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Xuelu Chen, Ziniu Hu, and Yizhou Sun. 2022. Fuzzy logic based logical query answering on knowledge graphs.

Shib Dasgupta, Michael Boratko, Siddhartha Mishra, Shriya Atmakuri, Dhruvesh Patel, Xiang Li, and Andrew McCallum. 2022a. Word2Box: Capturing set-theoretic semantics of words using box embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2263–2276, Dublin, Ireland. Association for Computational Linguistics.

Shib Sankar Dasgupta, Michael Boratko, Siddhartha Mishra, Shriya Atmakuri, Dhruvesh Patel, Xiang Lorraine Li, and Andrew McCallum. 2022b. Word2box: Capturing set-theoretic semantics of words using box embeddings.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Gerald Folland. 1999. *Real analysis : modern techniques and their applications*, volume 1. Wiley.

William L. Hamilton, Payal Bajaj, Marinka Zitnik, Dan Jurafsky, and Jure Leskovec. 2019. Embedding logical queries on knowledge graphs.

Zijie Huang, Daheng Wang, Binxuan Huang, Chenwei Zhang, Jingbo Shang, Yan Liang, Zhengyang Wang, Xian Li, Christos Faloutsos, Yizhou Sun, and Wei Wang. 2023. Concept2Box: Joint geometric embeddings for learning two-view knowledge graphs. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10105–10118.

Song Jiang, Qiyue Yao, Qifan Wang, and Yizhou Sun. 2023. A single vector is not enough: Taxonomy expansion via box embeddings. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 2467–2476, New York, NY, USA. Association for Computing Machinery.

Wei Ju, Zheng Fang, Yiyang Gu, Zequn Liu, Qingqing Long, Ziyue Qiao, Yifang Qin, Jianhao Shen, Fang Sun, Zhiping Xiao, et al. 2023. A comprehensive survey on deep graph representation learning. *arXiv preprint arXiv:2304.05055*.

E.P. Klement, R. Mesiar, and E. Pap. 2013. *Triangular Norms*. Trends in Logic. Springer Netherlands.

Chen Liang, Wenguan Wang, Jiaxu Miao, and Yi Yang. 2023. Logic-induced diagnostic reasoning for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16197–16208.

Junyu Lu, Bo Xu, Xiaokun Zhang, Changrong Min, Liang Yang, and Hongfei Lin. 2023. Facilitating fine-grained detection of chinese toxic language: Hierarchical taxonomy, resources, and benchmarks. *arXiv preprint arXiv:2305.04446*.

Shib Sankar Dasgupta Michael Boratko, Dhruvesh Patel and Andrew McCallum. 2022. Measure-theoretic set representation learning.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Steven Nahmias. 1978. Fuzzy variables. *Fuzzy Sets and Systems*, 1(2):97–110.

Alexander Panchenko, Stefano Faralli, Eugen Ruppert, Steffen Remus, Hubert Naets, Cédrick Fairon, Simone Paolo Ponzetto, and Chris Biemann. 2016. TAXI at SemEval-2016 task 13: a taxonomy induction method based on lexico-syntactic patterns, substrings and focused crawling. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1320–1327, San Diego, California. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Hongyu Ren, Weihua Hu, and Jure Leskovec. 2020. Query2box: Reasoning over knowledge graphs in vector space using box embeddings.

Hongyu Ren and Jure Leskovec. 2020. Beta embeddings for multi-hop logical reasoning in knowledge graphs.

Ryan A Rossi, Nesreen K Ahmed, Eunyee Koh, Sungchul Kim, Anup Rao, and Yasin Abbasi-Yadkori. 2020. A structural graph representation learning framework. In *Proceedings of the 13th international conference on web search and data mining*, pages 483–491.

Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang, and Jiawei Han. 2020. TaxoExpan: Self-supervised taxonomy expansion with position-enhanced graph neural network. In *Proceedings of The Web Conference 2020*. ACM.

Vered Shwartz, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2389–2398, Berlin, Germany. Association for Computational Linguistics.

Rion Snow, Daniel Jurafsky, and Andrew Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *Advances in Neural Information Processing Systems*, volume 17. MIT Press.

Emile van Krieken, Erman Acar, and Frank van Harmelen. 2022. Analyzing differentiable fuzzy logic operators. *Artificial Intelligence*, 302:103602.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Benedikt Wagner and Artur S d'Avila Garcez. 2022. Neural-symbolic reasoning under open-world and closed-world assumptions. In *CEUR Workshop Proceedings*, volume 3121. CEUR.

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, Ying Wang, and Yi Chang. 2021. Structure-augmented text representation learning for efficient knowledge graph completion. In *Proceedings of the Web Conference 2021*, pages 1737–1748.

David R. Wilkins. 2016. The multidimensional riemann-darboux integral.

Zhibiao Wu and Martha Palmer. 1994. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 133–138, Las Cruces, New Mexico, USA. Association for Computational Linguistics.

Hongyuan Xu, Ciyi Liu, Yuhang Niu, Yunong Chen, Xiangrui Cai, Yanlong Wen, and Xiaojie Yuan. 2023. Tacoprompt: A collaborative multi-task prompt learning method for self-supervised taxonomy completion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15804–15817.

Zezhong Xu, Wen Zhang, Peng Ye, Hui Chen, and Huajun Chen. 2022. Neural-symbolic entangled framework for complex query answering. *Advances in Neural Information Processing Systems*, 35:1806–1819.

Dongran Yu, Bo Yang, Qianhao Wei, Anchen Li, and Shirui Pan. 2022. A probabilistic graphical model based on neural-symbolic reasoning for visual relationship detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10609–10618.

Xinchen Yu, Ashley Zhao, Eduardo Blanco, and Lingzi Hong. 2023. A fine-grained taxonomy of replies to hate speech. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7275–7289.

Yue Yu, Yinghao Li, Jiaming Shen, Hao Feng, Jimeng Sun, and Chao Zhang. 2020. STEAM: Self-supervised taxonomy expansion with mini-paths. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery &amp Data Mining*. ACM.

L.A Zadeh. 1978. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 1(1):3–28.

L.A. Zadeh. 1999. Fuzzy sets as a basis for a theory of possibility. *Fuzzy Sets and Systems*, 100:9–34.

Rui Zhang, Bayu Distiawan Trisedya, Miao Li, Yong Jiang, and Jianzhong Qi. 2022. A benchmark and comprehensive survey on knowledge graph entity alignment via representation learning. *The VLDB Journal*, 31(5):1143–1168.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, Hua Jin, and Dacheng Tao. 2023. Knowledge graph augmented network towards multiview representation learning for aspect-based sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*.

Zhaocheng Zhu, Mikhail Galkin, Zuobai Zhang, and Jian Tang. 2022. Neural-symbolic models for logical queries on knowledge graphs.

A List of Symbols and Notations

In table 5, we list all the symbols and notations used in the paper, then we provide the theoretical proofs for the main results in the paper.

B Formal Statement of the Compactness Assumption

Here we state the assumption regarding the universe of discourse formally:

Assumption 1. *The universe of discourse U is topologically compact and has an open cover.*

Assumption 2. *The universe of discourse U is measurable and is associated with a measure space $(U, \mathcal{F}, \xi)$, where $\mathcal{F}$ is the σ-algebra and ξ its associated σ-finite measure. Moreover, $\forall A \in \mathcal{F}$, the fuzzy membership function $m_A : U \rightarrow [0, 1]$ is ξ-measurable.*

C Proof of Main Results

In this part we provide the proof sketches of the main results:

Lemma 1: *Let U be the universe of discourse and A a fuzzy subset of U with continuous membership function m_A, and let $\left(\mu_A^{(t)}\right)$ be a sequence of simple membership functions of m_A in Definition 2, such that $\{U_i\}_{i=1}^{n_t}$ is a refinement of $\{U_i\}_{i=1}^{n_{t-1}}$ ($n_t > n_{t-1}$), when t goes to infinity, then $\mu_A^{(t)}$ converges to m_A in the point-wise sense.*

Proof. The proof takes 2 steps:

Claim 1: By construction, the sequence $\left(\mu_A^{(t)}\right)$ is a sequence of monotonic non-increasing function and it is bounded below by the fuzzy membership function m_A.

Proof. The case where U is finite or countably infinite is straightforward. For the case where cardinality of U is uncountable but U is compact, we have the following: By definition 2, we have that for $i \in \{1, \cdots, d\}$ and $n \in \mathbb{N}$, $\forall u \in U_i$, $\mu_A^{(t)}(u) = \sup_{u \in U_i} m_A(u) \geq m_A(u)$. Hence on the entire domain, $\mu_A^{(t)}(u) \geq m_A(u)$. So $\forall t \in \mathbb{N}^+$, $m_A^{(t)}$ is an upperbound for m_A. Since $n_t > n_{t-1}$ indicates $\{U_i\}_{i=1}^{n_t}$ is a finer-grained partition than $\{U_i\}_{i=1}^{n_{t-1}}$, and the fact that supremum of a function

Symbol	Description
$A, B, C, \cdots$	mathematical sets.
$e_c, e_p, \cdots$	entities in a taxonomy
$E_c, E_p, \cdots$	fuzzy sets associated with entities
U	The universe of discourse, the set of all concepts.
$\mathbb{N}^+$	The set of all positive integers.
$(U, \mathcal{F}, \xi)$	A measure space with universe U, sigma-algebra $\mathcal{F}$ and a measure ξ.
A^c	The complement of a set.
$A \cap B$	Intersection of two sets.
$A \cup B$	Union of two sets.
(A, m_A)	a fuzzy set, where $A \subset U$, $\mathrm{m}_A : U \rightarrow [0, 1]$.
m_A	The fuzzy membership function associated with a fuzzy set.
(A, μ_A)	a simple fuzzy set, where $A \subset U$, $\mu_A : U \rightarrow [0, 1]$.
μ_A	The simple membership function associated with a simple fuzzy set (A, μ_A).
$(\mu_{A,n})$	a sequence of simple membership functions with monotonically finer-grained partitions.
$\mathcal{U}_A$	The fuzzy set embedding of (A, m_A).
$\mathbb{P}(A)$	The fuzzy measure (volume) of a fuzzy set under some measure space $(U, \mathcal{F}, \xi)$.
$\mathbb{P}_\mu(A)$	The simple fuzzy measure (volume) of a fuzzy set embedding under some measure space $(U, \mathcal{F}, \xi)$.
$\mathcal{M}$	Fuzzy mapping, which maps an input element into its associated Fuzzy Set Embedding.
P	A probability measure defined on the space of concepts.

Table 5: Table for all the symbols used in this paper

over finer grained partition is not greater than supremum over coarse-grained partition (e.g., supremum is monotonic w.r.t partitions), we have a non-increasing sequence of functions. □

Claim 2: The sequence of functions $\left(\mu_A^{(t)}\right)$ converges to m_A.

Proof. Since the space U is compact, we can define the space of function to be a compact Banach space of functions $f : U \to [0, 1]$ and so m_A is in the space. By the monotone convergence theorem (Folland, 1999) and by the fact that supremum over a singleton set $\{x\}$ is simply $\sup_{u\in\{x\}} m_A(u) = m_A(x)$, we can conclude that convergence holds in the point-wise sense. □

□

Theorem 1: *Let U be a compact universe of discourse and $(U, \mathcal{F}, \xi)$ a measure space. Let A be a fuzzy subset of U and m_A its membership function that's measurable. Moreover, let μ_A be its fuzzy set embedding membership function, then $\forall \epsilon > 0$, $\exists \delta > 0, d > 0$ such that if $d\delta = \xi(U)$ and $||U_i|| := \min_i \xi(U_i) < \delta$, we have:*

$$0 < \mathbb{P}_\mu(A) - \mathbb{P}(A) < \epsilon.$$

Proof. By Lemma 1 and assumption that m_A and hence all $\{\mu_A^{(t)}\}$ are ξ-measurable, we have a monotonic non-decreasing sequence of non-negative simple functions that converge point-wise to m_A. Then by the Monotone Convergence Theorem of simple functions (Folland, 1999), we have that $\int_U m_A d\xi = \lim_{t\to\infty} \int \mu_A^{(t)} d\xi$. Moreover, since $\mu_A^{(t)} \geq m_A, \forall n \in \mathbb{N}^+$, we have that $\forall \epsilon > 0, \exists N \in \mathbb{N}$, such that $\forall t > N, \mathbb{P}_\mu(A) > \mathbb{P}(A) < \epsilon$. This is equivalent to say (by virtue of construction of simple functions in definition 2) that $\exists \delta > 0, d > 0$, such that $d\delta = \xi(U)$ and $||U_i|| := \min_i \xi(U_i) < \delta$ (or equivalently, the partition is sufficiently fine-grained), the conclusion holds. □

Here we also provide a proof sketch for the Euclidean case, where the universe U is mapped into a compact subspace $\mathcal{U} \subset \mathbb{R}^d$, and the fuzzy measure is defined as a Riemann integral in $\mathbb{R}^d$ (the Euclidean volume), which is often the case for the embedding space. In this case, the following formulation of the theorem holds:

Theorem 1 (Euclidean Case) *Let $\Omega \subset \mathbb{R}^d$ be compact and let (A, m_A) be a fuzzy set with membership function $m_A : \mathcal{U} \to [0, 1]$, and let $\Omega = \bigcup_{i=1}^d \mathcal{U}_i$ be a partition and (A, μ_A) its associated partition-level fuzzy set. Then if m_A is Riemann-integrable on Ω, then $\forall \epsilon > 0$, $\exists \delta > 0, d > 0$ such that if $d\delta =$ Vol(Ω) and $||\mathcal{U}_i|| := \min_i$ Vol$(\mathcal{U}_i) < \delta$, we have that:*

$$0 < \mathbb{P}_\mu(A) - \mathbb{P}(A) < \epsilon$$

that is, the possibility of a partition-level fuzzy set is an upperbound for its underlying fuzzy set and it converges to the possibility of its underlying fuzzy set when the partition is sufficiently fine-grained.

Proof. The proof takes 3 main steps: Since Ω and $[0, 1]$ are both compact, we need to show that (a) under a rectangular partitions in $\mathbb{R}^d$, as partition granularity increases, the possibility $\mathbb{P}_\mu(A)$ monotonically decreases and (b) the possibility $\mathbb{P}_\mu(A)$ is an upper bound of the possibility $\mathbb{P}(A)$. After these two are shown, we can simply invoke the standard Monotone Convergence Theorem for compact spaces and show that $\mathbb{P}_\mu(A)$ converges to $\mathbb{P}(A)$ (Folland, 1999; Wilkins, 2016).

Step 1: *Show that $\mathbb{P}_\mu(A)$ monotonically decreases under granular partition*. This is equivalent to show that the upper Darboux sum of a $d-$dimensional Riemann integral monotonically decreases as the partition get finer-grained. This result is Lemma 6.4 in (Wilkins, 2016).

Step 2: *Show that $\mathbb{P}_\mu(A) \geq \mathbb{P}(A)$*. This is equivalent to say that $d-$dimensional upper Darboux sum is an upperbound for its Darboux-Riemann integral. This result is Lemma 6.6 in (Wilkins, 2016).

Step 3: *Show that $\mathbb{P}_\mu(A)$ converges to $\mathbb{P}(A)$*. By step 1 and step 2, we can construct a sequence of monotonically decreasing upper Darboux sums. By the Riemann-integrability of the function m_A and the compactness of the set $\Omega, [0, 1]$, we can conclude that by Monotone Convergence Theorem (Folland, 1999), this conclusion holds. □

Corollary 1 *Following the condition in definition 4, if in addition m_A is Lipschitz-continuous, then the convergence rate in 1 is $O(1/n)$. If m_A instead has bounded variation, then the convergence rate is also $O(1/n)$, where n is the number of partitions.*

Proof. In our case we defined the simple fuzzy set membership function as:

$$f_n(x) = \sum_{i=1}^{n} \sup f(x) \mathbf{1}(x \in U_i)$$

where $(f_n)_{n\in\mathcal{N}}, f$ are measurable functions on the space $(U, \mathcal{F}, \mu)$ and $\mathbf{1}$ is the indicator function, and we have $U = \bigcup_{i=1}^{n} U_i$ a partition of the universe. To derive a bound for convergence rate, we need to evaluate $\int |f_n - f| d\mu = \int (f_n - f) d\mu$, since $f_n \geq f$. and we have the following result:

$$\begin{aligned}\int (f_n - f) d\mu &= \sum_{i=1}^{n} \sup f(x) \mathbf{1}(x \in U_i) - \sum_{i=1}^{n} \int_{U_i} f d\mu \\ &= \sum_{i=1}^{n} \int_{U_i} \big(\sup_{x \in U_i} f(x) - f(x) \big) d\mu\end{aligned}$$

Now suppose that f is Lipschitz continuous with constant L (a much stronger assumption), then we have that

$$\forall x \in U_i, | \sup_{x \in U_i} f(x) - f(x)| \leq L\mu(U_i)$$

Then we have

$$\begin{aligned}\int (f_n - f) d\mu &= \sum_{i=1}^{n} \int_{U_i} \big(\sup_{x \in U_i} f(x) - f(x) \big) d\mu \\ &\leq \sum_{i=1}^{n} \int_{U_i} L\mu(U_i) d\mu = \sum_{i=1}^{n} L\mu(U_i)^2\end{aligned}$$

without loss of generality, let $\{U_i\}$ be an even partition and let $\mu(U) = 1$, then we have that $\mu(U_i) \propto O(\frac{1}{n})$ and so $\mu(U_i)^2 \propto O(1/n^2)$. Hence Equation 6 decays with $O(1/n)$ rate.

Suppose that f has bounded total variation $V(f) < \infty$, then we have that:

$$\begin{aligned}\int (f_n - fd) \mu &= \sum_{i=1}^{n} \int_{U_i} \big(\sup_{x \in U_i} f(x) - f(x) \big) d\mu \\ &\leq \sum_{i=1}^{n} \frac{V(f)}{n} \mu(U_i)\end{aligned}$$

Again, we have the conclusion that the error decays with $O(1/n)$.

□

D Details on Experiments

D.1 Baselines

For our taxonomy expansion task, we compare with existing methods that use vector embeddings or geometric embeddings. For vector embedding methods, we include also models that use advanced structures of taxonomy data. To summarize, the baselines we compare with are:

- **TAXI**(Panchenko et al., 2016): This is a vector-based embedding model that relies heavily on hypernym and hyponym relations between entities.
- **HypeNet**(Shwartz et al., 2016): This is a vector-based embedding model that leverages dependency paths between entity pairs.
- **Bert+MLP**(Yu et al., 2020): This is a vector-based embedding method that uses Bert (Devlin et al., 2019) to generate entity embeddings. Bert used in this model and in our own model is fine-tuned with a smaller learning.
- **TaxoExpan**(Shen et al., 2020): This is a vector-based embedding method leverages local ego-graphs to model pair dependencies. It uses graph neural networks (GNN).
- **STEAM**(Yu et al., 2020): This is a vector-based embedding method that samples dependency paths from taxonomy for better structured entity embedding.
- **BoxTaxo**(Jiang et al., 2023): This is a geometric embedding method that uses box embedding for entities in taxonomy. It is able to capture asymmetric relationships between entities.

D.2 Evaluation Metrics

For evaluating the taxonomy expansion task, we follow (Jiang et al., 2023). For the i-th query, denote a_i the true anchor and $\hat{a}_i$ the top-1 predicted anchor and let N be the total number of test samples, then the three metrics we use are the following:

- **Accuracy (ACC)**: evaluates the prediction's overall correctness

$$ACC = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{a}_i = a_i)$$

- **Mean Reciprocal Rank (MRR)**: evaluate the rank of the correct prediction in all predictions

$$MRR = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{rank(a_i)}$$

- **Wu & Palmer similarity (Wu& P)** (Wu and Palmer, 1994): measures the semantic similarity between concepts in a taxonomy

$$Wu\&P = \frac{1}{N} \sum_{i=1}^{N} \frac{2 \times \text{depth}(\text{LCA}(\hat{a}_i, a_i))}{\text{depth}(\hat{a}_i) + \text{depth}(a_i)}$$

where LCA is the least common ancestor of two inputs and depth is the depth in the taxonomy tree.

D.3 Implementation Detail for Base Model

In figure 5 and 6, the base model Fuzzy Set Embedding (FUSE) is the configuration of model with un-normalized global weights corresponding to volumes of each partition. In later ablation studies, we examine the impact of normalization on the volumes, corresponding to a choice of different measure. The number of partition of FUSE on the science dataset is 500, while the number of partition used on the environment dataset is 350. For training stability, we also normalize the fuzzy set embedding by their Euclidean norm before multiplication with volume weights. In addition, to make a fair comparison against baselines, we use the same optimization setup as in (Jiang et al., 2023) and provide a version of FUSE (FUSE ($\lambda = 1.0$)) with equal weights on intersection and asymmetry loss.

E Scope and Limitation

This work is the first attempt to use fuzzy set to model concepts in taxonomy expansion. To examine only the effectiveness of fuzzy set representation, we only use simple neural architectures and use only the child-parent pairs. The full capacity of FUSE should be further examined using datasets that contain First Order Logic (FOL) statements, since by construction, fuzzy sets should satisfy all the fuzzy logic axioms (Chen et al., 2022). This suggests future directions to expand taxonomy datasets with more complicated queries, and to handle more graph-structured data in social analysis and text mining (Ren and Leskovec, 2020; Ren et al., 2020; Chen et al., 2022; Zhu et al., 2022; Ju et al., 2023). Moreover, we can explore more explicit form of fuzzy membership function, such as a mixture of Gumbel boxes, to make the learning more concrete.