Taxonomy Embeddings on PubMed Article Subject Headings

Alyssa Lees, Jacek Korycki, Chris Welty, and Shubin Zhao

Google Research, USA
{alyssalees,korycki,shubin}@google.com, cawelty@gmail.com

Abstract. Machine learning approaches for hierarchical partial-orders, such as taxonomies, are of increasing interest in the research community, though practical applications have not yet emerged. The basic intuition of hierarchical embeddings is that some signal from taxonomic knowledge can be harnessed in broader machine learning problems; when we learn similarity of words using word embeddings, the similarity of *lion* and *tiger* are indistinguishable from the similarity of *lion* and *animal*. The ability to tease apart these two kinds of similarities in a machine learning setting yields improvements in quality as well as enabling the exploitation of the numerous human-curated taxonomies available across domains, while at the same time improving upon known taxonomic organization problems, such as partial or conditional membership. We explore some of the practical problems in learning taxonomies using Bayesian Networks, partial order embeddings, and box lattice embeddings, where box containment represents category containment. Using open data from PubMed articles with human assigned MeSH labels, we investigate the impact of taxonomic information, negative sampling, instance sampling, and objective functions to improve performance on the taxonomy learning problem. We discovered a particular problem for learning box embeddings for taxonomies we called the box crossing problem, and developed strategies to overcome it. Finally we make some initial contributions to using taxonomy embeddings to improve another learning problem: inferring disease (anatomical) locations from their use as subject labels in journal articles. In most experiments, after our improvements to box models, the box models outperformed the simpler Bayes Net approach as well as order embeddings.

Keywords: box embeddings · taxonomies · semantic relations

1 Introduction

Recent work in machine learning on representational structures such as taxonomies promises to blend the value of human curated taxonomies with the power and flexibility of machine learning systems. There are many human curated taxonomies such as in libraries, medicine and popular culture. They are a widely used organizational tool for inventories and information in general. Yet, they rely on an assumption of discreteness – category membership is either true or false – and this assumption does not generally play well with continuous valued systems or spatial embeddings.

Much of the recent machine learning work on taxonomies has focused on the toy problem of reconstructing an existing taxonomy, which is of little practical value except as a stepping stone to a broader interaction. Rather, the key question is whether taxonomic similarity can be teased apart from others. For example, we want to say that while lion, tiger, and animal are all similar, animal is more general. In many cases, this explicit knowledge can be derived from existing taxonomies, but without some way of understanding this knowledge in the embedding space, it is just another similarity signal. Recently, order embeddings addressed this issue by orienting similarity in two axes and adding a constraint to similarity embeddings that ensured more general terms had to be closer to the origin.

More recently we've seen the introduction of box lattice embeddings, which treat categories in a taxonomy as n-dimensional boxes, with a constraint that boxes representing more general terms should contain boxes representing more specific terms.

In this work, we are primarily interested in learning a typical inference problem in which there are two categories of entities with a target of learning the relationships between them. We have chosen an open dataset to exemplify the problem: PubMed articles with human labeled subject headings. We want to learn the relationship between diseases and parts of the body using the co-occurrence of these labels as subjects and further informed by the inclusion of taxonomic information. It stands to reason that if asthma (disease) and bronchi (anatomy) co-occur in data, and bronchi is a subcategory of lung in the taxonomy, then asthma and lung are related. Our primary research question is whether these two types of knowledge (co-occurrence and taxonomy) can be integrated in a way that can capture softer constraints than, for example, a set of logic rules. However, in order to answer this question we had to explore the practicalities of learning embeddings for categories in a taxonomic structure.

The contributions of this paper center on rigorous analysis of taxonomy embeddings from data, in particular training regimes that impact the faithfulness of the embeddings, and the degree to which taxonomic information can be incorporated in a learning problem. Our experiments focused on Box Lattice models, though we also include some analysis of Bayes Nets and Partial Order Embeddings. Our first contribution is a simple fix to the non-differentiable hinge property of the original box model objective function. We then examine the influence of negative sampling methods and weighting on learning the taxonomy. We argue that, while sensitivity to negatives is not a particularly new machine learning problem, the peculiar nature of learning boxes with volumes – as opposed to vectors – makes this increasingly important. Next we consider various options for using instance data to support the taxonomy embedding task, and contribute methods and analysis for summarizing instances to avoid explosion of the parameter space. We also perform analysis of metrics for evaluation, noting that the generalization implicit in taxonomies makes some categories easier and others harder.

2 Background

MeSH, the NLM *Medical Subject Headings*, [1] is a taxonomy of subject headings for categorizing medical writing. It is organized in a traditional library subject classification style, with the slight difference that subject headings in the taxonomy can have multiple parents – it is a DAG not a tree. Pubmed [2] is a very large collection of meta-data for over 30 million medical journal articles, each with human labeled subject categories from MeSH. The medical journal community of authors, editors, publishers and the NLM's librarians, work together to keep the process scalable and of reasonable quality. That's not to say there aren't errors, but no taxonomy is free of errors.

MeSH is organized into 16 top level categories, such as `A: Anatomy, B: Organisms, C: Diseases`, which themselves cannot be the subjects of articles. The taxonomy is on average 8 levels deep, with a generalization or broader-term relationship from child to parent nodes, e.g. `<Respiratory System, Anatomy>, <Larynx, Respiratory System>, <Asthma, Diseases>`. For simplicity we will adopt this notation, where the taxonomy is represented linearly as a collection of $\langle child, parent \rangle$ *edges* in the tree-shaped taxonomy graph (noting again that it is not strictly a tree).

The MeSH anatomy hierarchy mostly follows a Mereological generalization (subparts to parts), while diseases follow a slightly more causal generalization (`<Pneumonia, bacterial Infection>`). This kind of semantic promiscuity is extremely common in taxonomies [5], and causes an imprecision that begs for an approach with soft, continuous-valued constraints, as opposed to the traditional discrete reading of the taxonomic relationship (all members of the subcategory are members of the super-category). It was this observation that led us to taxonomy embeddings.

Articles listed in PubMed can have any number of subject headings, on average 8-10, and it is fully expected by the published methodology that assigning a particular subject

heading to an article *also assigns the MeSH parents and ancestors* – they expect the transitive closure to hold.

Most taxonomies take a universal vs. particular (ie category vs. instance) view of the world, that instances are members of set-like categories. The primary semantic distinction is that the relationship between instances and categories (instance-of) is not transitive, whereas the relationship between categories (subcategory) is transitive. However, like the subcategory relations shown above, the instance-of relation is very promiscuous in practice, and we shall continue that practice by considering the relationship between articles in PubMed and MeSH categories as the instance-of relation, the articles are *members of* their subject categories.

We side-step whatever philosophical fallout these choices may entail by defining our research question to be a data-mining one:

> **RQ1**: Can we use the co-occurrence of subject labels on PubMed articles to learn associations between those labels, similar to [10]? As a simplifying example, can we learn the association between diseases and parts of the body from the co-occurrence of MeSH labels on PubMed articles in the Anatomy and Diseases branch of the MeSH taxonomy?
>
> **RQ2**: Can the taxonomy be used to improve the learning of these associations? Intuitively, knowing `<asthma, lung disease>` and `<bronchi, lung>` should help us associate asthma to lung and even lung disease to lung.

The principal contributions of this paper are an empirical analysis of these embedding techniques, primarily box embeddings, and how the basics of the taxonomy problem interact with learning.

3 Taxonomy Embeddings

3.1 Partial Order Embeddings

We applied Partial Order Embeddings to our datasets, an established technique for modeling taxonomy data, as described in [11]. The model assigns each entity u an embedding $f(u) \in \Re^n$ in such a way that the order in the space of entities, defined by edges (u, v), is maintained in the embedding space by requiring that $f(u) \geq f(v)$, which holds for all dimensions. This is accomplished by defining a continuous score function for embeddings that measures compliance with the order constraint:

$$E(x, y) = ||max(0, y - x)||^2$$

This score should be zero, or close to it, for a pair forming an edge (from a set of positive examples P) and large positive for a pair that is not an edge (from a set of negative examples N). This requirement is enforced by minimizing the following max-margin loss function over embeddings f:

$$\sum_{(u,v)\in P} w(u, v)E(f(u), f(v)) + W \sum_{(u',v')\in N} max(0, \alpha - E(f(u'), f(v')))$$

It uses a margin $\alpha > 0$ to express the minimum desired value for a score of the negative example. Relative importance of positive examples may be controlled with edge weights w, if such values are available in the dataset (for example as conditional probabilities). Hyper-parameter W can be used to control relative importance of positive and negative parts of the loss.

3.2 Box Lattice Embeddings

Box lattice embeddings associate each category with 2 vectors in [0, 1], (x_m, x_M), the minimum and maximum value of the box at each dimension [12]. For numerical reasons these are stored as a minimum, a positive offset plus an ϵ term to prevent boxes from becoming too small. This representation of boxes in Cartesian space can define a partial ordering by containment of boxes, and a lattice structure as:

$x \wedge y = \perp$ if x and y disjoint, else
$x \wedge y = \prod_i [max(x_{m,i}, y_{m,i}), min(x_{M,i}, y_{M,i})]$
$x \vee y = \prod_i [min(x_{m,i}, y_{m,i}), max(x_{M,i}, y_{M,i})]$
if $A = (x_m, x_M)$ then $p(A) = \prod_i^n (x_{M,i} - x_{m,i})$

and the objective function follows as

$$-log(p(a \vee b) - p(a) - p(b))$$

In other words, maximize the overlap of boxes with a positive edge. The original boxes paper showed results for reconstructing a taxonomy derived from WordNet using the transitive closure of edges, and the team made the code available in Github, which we reused and modified for our experiments.

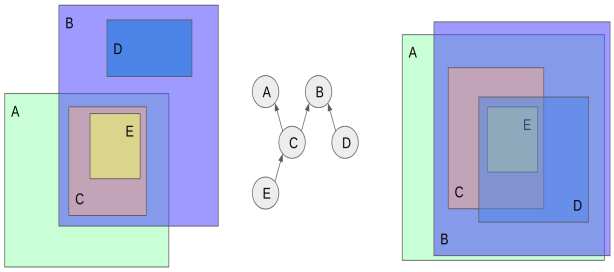

Fig. 1. An idealized box rendering of a simple taxonomy, and the actual learned box embedding. With no negative training data, all boxes simply overlap.

3.3 Bayes Nets

For comparison, we used a Bayesian model to predict whether an instance belongs to a category. Given the training data, the model computes the conditional probability $p(C_x|C_y)$ for any category pair (C_x, C_y) that exists in the training data. Negative training examples are ignored because the category pairs are randomly generated. There is no correlation between the categories. Given a test example (C, C'), the model needs to predict whether there is a positive relationship. (C, C') should not appear in the training data, but C and C' must appear in the training data. Otherwise we would have no clue to predict the correlation. Assuming the seen parent categories of C in the training data are $(C1, ..., Cm)$, The model makes an estimate that

$$p(C'|C) = \sum_{k=1}^{m} p(C'|C_k, C)p(C_k|C) \approx \sum_{k=1}^{m} p(C'|C_k)p(C_k|C)$$

.

Given a score threshold T, if $p(C'|C) > T$, the model predicts a positive relationship for (C, C'), otherwise it predicts a negative relationship. The precision and recall were picked as the best ones in the precision-recall curve with different T values.

4 Learning Taxonomies

There are many taxonomies for organizing many kinds of information, and they can be differentiated in many ways. For our purposes, we focus on a few that became important in our experiments.

Many taxonomies are *extensional*: the categories organize sets of entities in some problem domain, such as movies, stores, restaurants, books, songs, etc. There is a

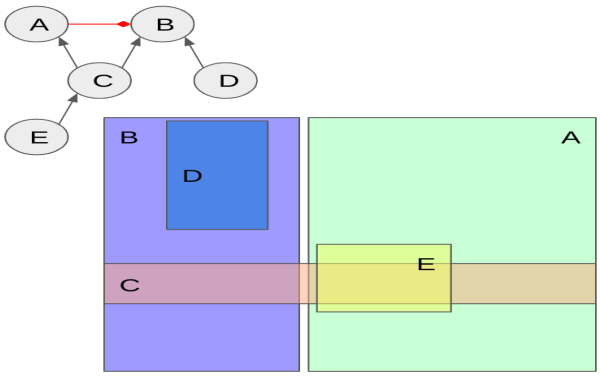

Fig. 2. The effect of a naive negative edge – the two root categories are forced apart, and category C must split its loss between the two.

fairly clear instance/category distinction in these cases and the number of instances vastly outweighs the number of categories. The set of instances of a category is its extension. Viewed in its entirety, such taxonomies have edges from instances to categories, which we call *inst-cat* edges (e.g. `<Star Wars, Science-Fiction Movie>`), and edges between categories, which we call *cat-cat* edges (e.g. `<Cult Science-Fiction Movie, Science-Fiction Movie>`).

Some taxonomies are *intensional*: they have no instances, at least none represented in data. WordNet synsets, for example, are arranged in a taxonomy, and there are very few instance-like synsets in WordNet, referring e.g. to specific people and organizations, but for the most part, synsets like "amount of matter" have no clear extension. Such taxonomies have only cat-cat edges.

Most previous experiments on learning taxonomy embeddings were done on intensional taxonomies, with no instances, and the objective (for training and evaluation) was simply the number of correct cat-cat edges learned above some confidence threshold.

Our datasets are extensional and contain millions of instances: PubMed 2018 has over 30 million. Surely this wealth of data could be used to help focus the taxonomy learning. At the very least, inst-cat edges could tell us the relative importance of getting certain cat-cat edges correct.

4.1 Learning taxonomies from instances

We tried many approaches to utilizing the rich extension of MeSH in PubMed articles. The most obvious was to treat each inst-cat edge as a part of the graph, and ignore the semantic differences with cat-cat edges. However, in our embedding techniques, edges are training examples and the vertices (e.g. each category or instance) are the learnable parameters, so this results in an explosion of parameters almost to the point of having fewer examples than parameters. Further, instances as embeddings causes many spurious inst-inst edges to be inferred from the dense encodings, generating tremendous noise. One solution may be to use different forms of optimization, such as annealing or Brownian Motion. We save this for future work.

Another approach is to reuse the embedding techniques designed for cat-cat edges and *summarize* the inst-cat edges into the categories. For every pair of co-occurring inst-cat edges$\langle c, p_1\rangle$ $\langle c, p_2\rangle$, we emit a cat-cat edge $\langle p_1, p_2\rangle$. This leads to repetition of edges in the training data that reflects the magnitude of the co-occurrence. We compared these two approaches:

Taxo: Use the edges from the taxonomy, with no consideration for the instances.

Summary: Include the co-occurrence of categories in instances as a weight on the taxonomy edges.

4.2 Transitive Closure

In addition to summarizing instances, we experimented with the use of deterministic reasoning on the training data, specifically the transitive closure of cat-cat and inst-cat edges. In previous work, this is the only approach taken. However it seems possible that, given only *direct* cat-cat edges (e.g. if we have $\langle a, b\rangle$ and $\langle b, c\rangle$, then we do not have $\langle a, c\rangle$), we can learn embeddings for the taxonomy that approximates it.

For instance edges, we can similarly compute the transitive closure in the usual way (e.g. if we have $\langle i, a\rangle$ and $\langle a, b\rangle$ then we add $\langle i, b\rangle$), and then summarize from the inferred inst-cat edges as described above.

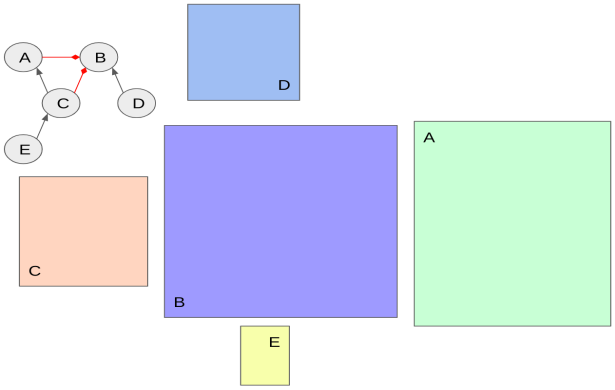

Fig. 3. The *box crossing* problem. Category C must cross the negative edge it has with B in order to get to A. In the original box model, box crossings cannot be overcome.

This gives us two more approaches for each data set:

Direct: do not compute the transitive closure of the category edges. For PubMed articles, only the human labelled subject heading are used. For the most part, this means no instance will have edges to a category and its ancestor, however there are a few exceptions. Reconstructing a taxonomy from only direct edges is expected to be an extremely hard problem to solve with embedding methods, but does serve the purpose of informing us how these particular methods can be used to learn other relations, and how to isolate the signal coming from the taxonomy itself. See more discussion below.

Closure: compute the transitive closure of the category edges. PubMed articles will have edges to each human supplied subject heading and all its ancestors. Clearly the closure multiplies the positive training data by large amount, and this is expected to be the simpler of the two approaches to learn the taxonomy.

4.3 Negative sampling

In the original box and order embedding papers, negatives were sampled uniformly from the complement of positive edges with a ratio of 1:1.

We found this uninformed sampling created conflicting constraints and could generate negatives that prevented desirable generalizations. Some improvements to sampling were proposed in [3], but these were more specific to order embeddings.

Clearly negative sampling is a general problem in machine learning, but again a bit of analysis and understanding of the taxonomy problem can help reduce the impact of the problem.

When we tried to replicate the original box embedding results on our taxonomy, we found that manual inspection of the untrained positives (i.e. the model generalizations) showed many bad edges, such as $\langle heart, lung\rangle$ or $\langle bacteria, virus\rangle$. As we added more test negatives, we found the false positive rate was simply a function of the neg:pos ratio. For boxes it is somewhat trivial with a low neg:pos ratio to draw large boxes that overlap

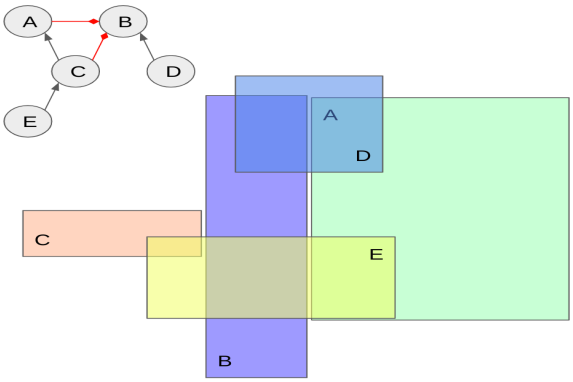

Fig. 4. With the box hinge loss fixed, the two negatives B and C thin themselves in preparation for crossing.

nearly everything, or long thin boxes that overlap with exactly the set of boxes they need to, while avoiding the few negatives. In Fig. 1, the idealized box embedding for a simple taxononmy is shown on the left; category C should be contained in both A and B, therefore the latter should overlap. Without negatives, the simplest zero-loss solution is to make all boxes overlap, as shown on the right of the figure. With a low negative rate (and many more boxes), the solutions ended up looking similar.

Raising the neg:pos ratio seems like an obvious solution, however at 10:1 box embeddings failed to learn anything on all our datasets, achieving 90% accuracy by making all boxes disjoint. We compared performance of the various approaches based on two neg:pos ratios:

neg1: Uses a 1:1 neg:pos ratio

neg10: Uses a 10:1 neg:pos ratio

In our error analysis we found negative edges that seemed to cause learning problems. First, the transitive closure of positive edges should not be in the the negative set. This is not a problem when the transitive closure is in the positive set, as with previous work, but the **direct** approach in Sec. 4.2) does not assume this, and so we must guard against it.

If negatives are generated on-the-fly (as they were in the original box embeddings implementation), negatives from the transitive closure can leak into training after a test/train split. We must generate negatives in batch and split into test/train.

Self edges should never be negative.

The inverse edges should never be negative, e.g. if $\langle c, p \rangle$ is a positive edge, $\langle p, c \rangle$ should not be negative. This is an interesting divergence between the box objective and the set semantics of categories: in a set semantics, the pair and its inverse can only be true if $c = p$, thus negating the inverse is equivalent to asserting strict subset. The box objective, however, pushes negative edges *apart*, and a sub-category must clearly overlap with its parent.

Edges between ancestors of the same category should never be negative, e.g. if $\langle c, p_1 \rangle, \langle c, p_2 \rangle$ are positive, then neither $\langle p_1, p_2 \rangle$ nor its inverse should be negative since, in a strict reading of taxonomy, their boxes should overlap. In Fig. 1, the idealized box embedding for a simple taxonomy is shown on the left; category C should be contained in both A and B, therefore the latter should overlap. Without negatives, the simplest zero-loss solution is to make all boxes overlap. In Fig. 1 the effect of a poorly chosen negative edge between A and B is shown, it forces the two boxes apart and C must deal with the loss. Again, in a set semantics, one may want to negate the subcategory edge between two overlapping categories, which would simply be interpreted as a constraint against one being a subset of the other. With boxes, doing so would again conflict with the objective to make them overlap.

We introduce *informed sampling* (Alg. 0) as a way of avoiding these problems. The algorithm incrementally builds $\mathcal{NN}$, the set of non-negative edges, by processing the

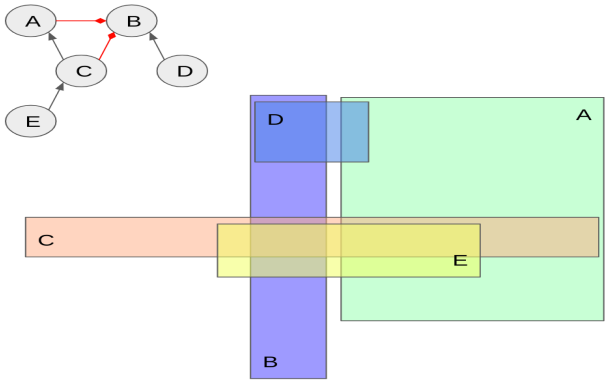

Fig. 5. With a lower weight on negatives, the crossing problem is solved by the thinned C piercing the thinned B to reach A.

constraints discussed above, and subtracts that from the edge set formed by the cross product of all categories in the taxonomy, $\mathcal{C}$.

Algorithm 1: Informed sampling of taxonomy negatives

Input: Set of positive edges $\mathcal{P} = \{\langle c, p\rangle, ...\}, c, p \in \mathcal{C}$
Output: Set of negative edges to sample from

$\mathcal{T} \leftarrow Closure(\mathcal{P})$
$\mathcal{NN} \leftarrow \mathcal{T}$
for $\langle c, p\rangle \in \mathcal{T}$ **do**
  $\mathcal{NN} \leftarrow \mathcal{NN} \cup \{\langle c, c\rangle, \langle p, p\rangle, \langle p, c\rangle\}$
  $map(c) \leftarrow map(c) \cup \{p\}$
for $c \in map$ **do**
  for $p_1 \in map(c)$ **do**
    for $p_2 \in map(c)$ **do**
      $\mathcal{NN} \leftarrow \mathcal{NN} \cup \{\langle p_1, p_2\rangle\}$
return $\mathcal{C} \times \mathcal{C} - \mathcal{NN}$

In experiments we compare these two approaches to generating negatives:

Naive sampling: Sample negatives from the complement of the transitive closure of positive edges.

Informed sampling: Sample negatives using informed sampling.

4.4 Box crossings

Despite discovering these problems in negative sampling, we were unable to get any experiments with boxes to learn anything with higher negative ratios.

In further analysis we found the cause to be what we named the *box crossing problem*, illustrated in Figs. 3, 4, 5. When two boxes representing categories that have a positive edge are on either side of a box for which they each have a negative edge, the cost for the two boxes to cross the negative one is too large. As the neg:pos ratio increases, the chances of box crossing standing in the way increases dramatically. While in some sense this is no more than a gradient descent problem, getting stuck in a local neighborhood, the specific learning problem, rather than some generic approach to optimization, gives us insight into solutions that smooth the objective.

In previous work, boxes were initialized with random positions. This feeds the box crossing problem; as the negative ratio increases, boxes are more likely to start off surrounded by negatives. To overcome this problem, we add two more experimental approaches:

Rand: Boxes have random initial size and position.

Center: Boxes all start out in a small central sphere position in the middle of the space. With all boxes equally overlapping positives and negatives, box crossing should not prevent them from moving in more optimal directions at the start.

In addition, we explored weighting the negative edges in training much lower than positives, to allow the positive signal to draw boxes together, enough to overcome box crossings, and once the positive constraints are satisfied, the boxes should then move and re-size to obey the negatives. This gives us another set of experimental approaches to explore:

nloss-n: The negative loss is weighted by n compared to the positive loss.

Although the failure of the original box model to generalize in the presence of more negatives forced us to discover and propose fixes to negative sampling, and the box crossing problem, ultimately we found the primary problem was a hinge property of the loss function. When the box crossing problem arises, there is a non-differentiable hinge in the negative loss on the step where the negative boxes first overlap. An approach to smoothing the loss has appeared very recently [8], we used a far simpler fix developed before that result came out, which in combination with lower negative weight is capable of overcoming the box crossing problem as shown in Fig. 5.

The hinge fix ensures a smooth transition when boxes with a positive edge meet, and when boxes with a negative edge move apart. When positive boxes are disjoint, there is loss proportional to their distance, when they meet the loss is inversely proportional to the amount of overlap. At the point where they meet, these two kinds of loss must also meet, otherwise an unbounded gradient occurs.

For negative edges, when two overlapping boxes come apart, the negative loss is zero. The original box model clipped this value at ϵ in order to avoid a zero value, and added simple smoothing. This was not working in practice, as we saw sharp gradient spikes during optimization. To solve this, we merely reversed the approach for positive loss, splitting the loss into overlapping and disjoint loss, and flipping the sense.

We do not present many experiments with the initial box models, as they were unable to learn anything with more than a 1:1 neg:pos ratio, for clarity we refer to the two versions as:

Box orig: the original box code without a hinge fix

Box: the box model with the hinge fix

5 Inter-taxonomy Relations

Our final set of experiments begins to touch on our central research questions (q.v. RQ1 & RQ2, page 3). In our experiments, we used several slices of data from MeSH [1] and PubMed [2]. Instances are summarized from the first ten shards of the PubMed 2018 database, consisting of 300k articles. We use the subject *descriptors*, which are associated with one or more coded entries in the strict MeSH taxonomy. Descriptors give MeSH a DAG structure, whereas the coded taxonomy is a tree.

Our datasets are grouped in three ways:

single taxonomy: we use one branch of the MeSH taxonomy, *C: Disease*, containing 4800 categories. In these experiments, like the previous literature, we attempt to reconstruct the taxonomy.

dual taxonomy: we use two branches of the MeSH taxononmy, *A: Anatomy* and *C: Disease*, for a total of 6610 categories. In these experiments, we attempt to learn the cross-taxonomy relations between diseases and their anatomical locations, while at the same time reconstructing each taxonomy. The intention is for the two taxonomies to overlay each other to represent the location relation. To measure the additional signal added by the taxonomy, we compare the direct and closure versions of the dual taxonomy datasets.

multi taxonomy: we use all branches of the MeSH taxonomy, a total of 28954 categories, summarized from 300k articles. In this case we simply try to reconstruct all edges, within and across taxonomies. We do not expect any technique to perform well here.

6 Experiments

In our first set of experiments, we attempt to reproduce the results of the original box embedding paper on a different taxonomy. We identified the following characteristics, described in Section 4, that we tested:

taxo vs. summary: Edges are straight from the taxonomy or summarized from instances

neg1 vs.neg10: neg:pos ratio is 1:1 or 10:1

direct vs. closure: Edges include the transitive closure of taxonomy or only direct edges

rand vs. center: Boxes are initialized in random position or at the center

naive vs. informed: Negatives are generated randomly or informed by the taxonomy

nloss1 vs. nloss.1: Negative loss is weighted or not

	Tags	box-orig	bayes	poe	box
1	taxo neg1 closure rand informed nloss1	0.33	0.77	0.89	0.91
2	taxo **neg10** closure rand informed nloss1	0.00	0.78	0.83	0.90
3	taxo neg10 closure rand **naive** nloss1	0.00	0.77	0.83	0.83
4	taxo neg10 closure **center** informed nloss1	N/A	N/A	N/A	0.91
5	taxo neg10 closure rand informed **nloss.1**	0.00	N/A	0.82	0.82
6	**summary** neg10 closure rand informed nloss.1	0.00	0.92	0.97	0.95
7	taxo neg10 **direct** rand informed nloss1	0.00	0.17	0.35	0.34
8	taxo neg10 direct rand **naive** nloss1	0.00	0.17		0.33

Table 1. F-measure results of single taxonomy experiments

The results are shown in Table 1. Experiment 1 replicates the original box embeddings paper using the code made available by that team. With 1:1 neg:pos sampling, as discussed above, the model shows non-zero performance on the test set, but on inspection this was due to over-generalization - many boxes overlapped many others on average. In experiment 2, we added a 10x as many negatives and the original code converges to an accuracy of 0.9 by making no boxes overlap. This is the baseline for the subsequent experiments.

Experiment 3 shows the impact of informed sampling of negatives. This does not affect POE or Bayes approaches, but has a large effect on boxes, dropping from .90 to .83. We illustrated the problem of naive negatives on boxes in Fig. 2.

Experiment 4 (compared to 2) is specific to boxes and shows a small increase in F-measure when boxes are initialized in the center of the space. While the increase is not significant, it does show a very good increase in speed of convergence of the models.

Experiment 5 shows that down-weighting negative instances does not have a positive effect on F-measure. In our early experiments, this did have a positive effect but appears to be overwhelmed by the other improvements. We intend to further vet this particular result as it contradicts results we saw in earlier experiments, based on which we ran the larger inter-taxonomy experiments with *nloss.1*.

In Experiment 6, we have replaced the taxonomy edges with edges resulting from the instance summarization process described in Sec. 4.1. This is the best peforming overall set of characteristics and seems to work well in combination with low negative weighting. It seems most likely this is due to the strategic repetition of edges, as opposed to simply repeating edges by cycling through a training set multiple times. Instance summarization identifies the important edges to re-sample. The technique helps POE and Bayes as well, giving POE the top single taxonomy score. In multi-taxonomy experiments below, instance summarization gives us the inter-taxonomy links as well, making this a powerful technique for taxonomy learning. It's worth noting that the BayesNet approach is incredibly simple, and that with instance summarization provides quite good performance.

Experiments 7 and 8 show the negative impact of dropping the transitive closure edges. As discussed in Sec. 4.2, it seems that simply having the direct links, e.g. $\langle c, b\rangle$, $\langle b, a\rangle$ and

not $\langle c, a \rangle$ should make it possible to learn that transitive edge, however our results show this simply doesn't happen in any approach. For boxes, the transitive edges provide a powerful amplification of the containment constraint, ensuring that when a higher level box moves, all its children move with it, as opposed to cascading that movement across the next several steps.

From manual inspection of the single taxonomy experiments, the results seem to hold up, and we do not find a lot of spurious edges. A deeper evaluation would be productive, however we will save that expense for future work on inter-taxonomy results.

Table 2 shows the results of our inter-taxonomy experiments. All were run with the winning set of *summary, neg10, closure, rand, informed*, and we included *nloss.1* as that showed better performance in previous tests. These are bigger experiments and take longer to validate, and will include a comparison of that feature in the final paper. For each set we compared direct and closure edges, continuing to confirm tha tthe closure edges are important for maintaining the taxonomy.

As mentioned above, we have not analyzed these results very deeply, the primary contributions of the paper are analysis of getting box embeddings to work on a single taxonomy. Like the original box embeddings results, the numbers for inter-taxonomy look outrageously good, however manual inspection does not tell the same story. It is likely we are still not covering the negative space adequately.

For example, while we do see edges associating diseases to parts of the body that make surprising sense, such as: `<Heart Septal Defects - Ventricular, Sarcoplasmic Reticulum>` which is not a terrible association, and a good generalization of that to `<Heart Valve Diseases, Sarcoplasmic Reticulum>`, we also see too many edges like `<Heart Failure, Temporal Bone>`. More work is needed.

		bayes	poe	box
1	dual direct	0.83	0.90	0.93
2	dual closure	0.85	0.96	0.98
3	multi direct	0.82	0.89	0.96
4	multi closure	0.80	0.91	0.97

Table 2. F-measure results of dual and multi taxonomy experiments

Finally, we explored depth-based scoring [13], on the intuition that higher level categories have more incoming edges (edges from their children), and are easier to classify as they are larger. However, we saw no difference when evaluating edges by their depth in the taxonomy, except in the cases where the closure edges were not in training.

7 Conclusion

We presented practical explorations of three types of embeddings for taxonomies: Bayes, Partial Order Embeddings, and Box Lattice Embeddings. We explored many characteristics of the learning problem of reconstructing a single taxonomy, with a focus in particular on the problem of generating appropriate negative examples, using signal from large instance-based taxonomies such as PubMed and MeSH. We discovered a particular problem for learning box embeddings for taxonomies we called the box crossing problem, and developed strategies to overcome it. We showed impressive F-measure scores for problems of reconstructing the MeSH Disease taxonomy from edges, as well as finding inter-taxonomy relations from associations between them in PubMed article subject headings, though these latter results have yet to be analyzed thoroughly. In most experiments, after our improvements to box models, the box models outperformed the simpler Bayes Net approach as well as Order Embeddings.

References

1. Mesh medical subject headings (2018), https://www.nlm.nih.gov/mesh/meshhome.html
2. Pubmed medical abstracts (2018), https://www.ncbi.nlm.nih.gov/pubmed/
3. Athiwaratkun, B., Wilson, A.G.: On modeling hierarchical data via probabilistic order embeddings. In: International Conference on Learning Representations (2018)
4. Globerson, A., Chechik, G., Pereira, F., Tishby, N.: Euclidean embedding of co-occurrence data. J. Mach. Learn. Res. **8**, 2265–2295 (Dec 2007)
5. Guarino, N., Welty, C.: An Overview of OntoClean, pp. 201–220 (05 2009)
6. Hoxha, J., Jiang, G., Weng, C.: Automated learning of domain taxonomies from text using background knowledge. Journal of Biomedical Informatics **63**, 295 – 306 (2016)
7. Langley, P.: Crafting papers on machine learning. In: Langley, P. (ed.) Proceedings of the 17th International Conference on Machine Learning (ICML 2000). pp. 1207–1216. Morgan Kaufmann, Stanford, CA (2000)
8. Li, X., Vilnis, L., Zhang, D., Boratko, M., McCallum, A.: Smoothing the geometry of probabilistic box embeddings. In: International Conference on Learning Representations (2019)
9. Liu, X., Song, Y., Liu, S., Wang, H.: Automatic taxonomy construction from keywords. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1433–1441. KDD '12 (2012)
10. Sarkar, P., Siddiqi, S., Gordon, G.: A latent space approach to dynamic embedding of co-occurrence data. In: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AI-STATS) (January 2007)
11. Vendrov, I., Kiros, R., Fidler, S., Urtasun, R.: Order-embeddings of images and language. ICLR (2016)
12. Vilnis, L., Li, X., Murty, S., McCallum, A.: Probabilistic embedding of knowledge graphs with box lattice measures. In: ACL (2018)
13. Yang, G.H., Callan, J.P.: A metric-based framework for automatic taxonomy induction. In: ACL/IJCNLP (2009)