huang_2020_corel_seed_guided_topical_taxonomy_construction_by_concept_learning_and_relation_transferring

Year

2020

Author(s)

Huang, Jiaxin and Xie, Yiqing and Meng, Yu and Zhang, Yunyi and Han, Jiawei

Title

CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring

Venue

KDD

Topic labeling

Partially automated

Focus

Primary

Type of contribution

Novel

Underlying technique

Transformer-based (weakly supervised relation classifier)
(As part of the proposed seed-guided topical taxonomy construction)

Topic labeling parameters

Relation classifier:

- batch size = 16
- training epochs = 5,
- model: Bert-Base (12 layers, 768 hidden size, 12 heads)
- 90/10 training/validation split on training samples.

In the relation transferring process:

- threshold for relation score (Equations 1 and 2): 0.7
- Threshold for KL divergence δ: 0.5.

Label generation

A relation transferring module learns the specific relation preserved in seed taxonomy and attaches new concepts to existing nodes to complete the taxonomy structure.

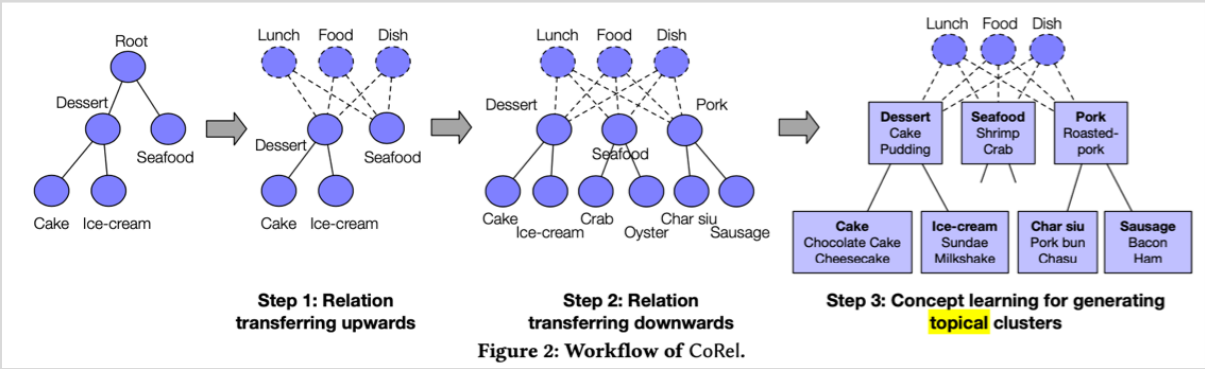

Figure 2: Workflow of CoRel.

The relation transferring module first captures the relation preserved in seed parent-child pairs and transfers it upwards and downwards for finding the first-layer topics and subtopics, attached by a co-clustering technique to remove inconsistent subtopics. The concept learning module learns a dis- criminative embedding space by jointly embedding the taxonomy with text and separating close concepts in the embedding space.

A pre-trained BERT model is used to train a relation classifier.

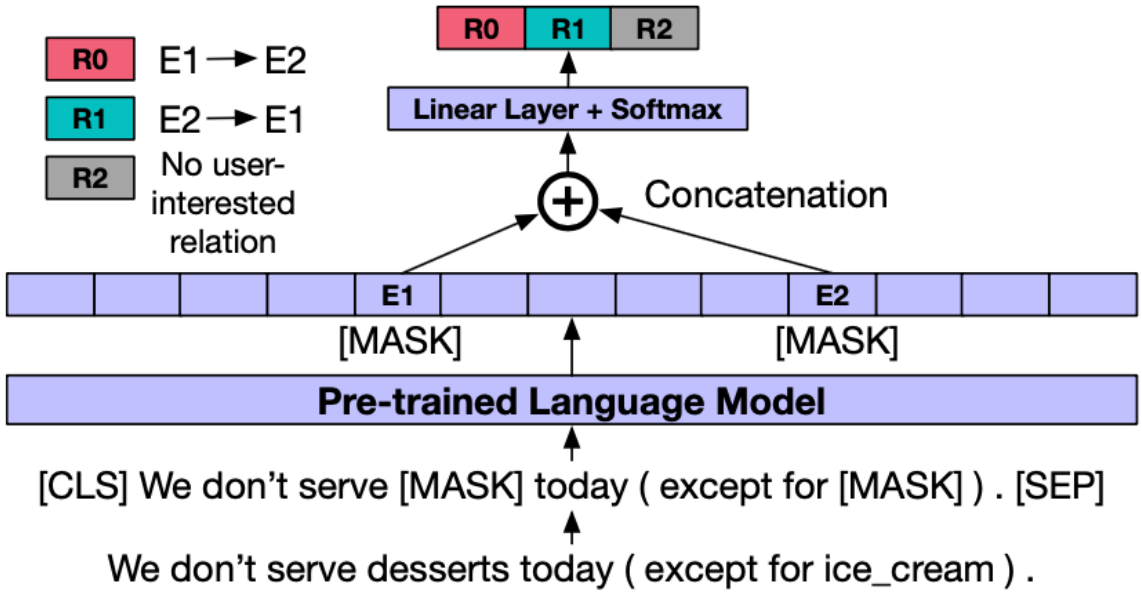

Figure 3: Our Weakly Supervised Relation Classifier.

Relation Statement

We assume that if a pair of ⟨p,c⟩ (parent and child nodes) co-occurs in a sentence in the corpus, then that sentence implies their relation.

We refer to sentences containing ⟨p,c⟩ as their relation statements and leverage BERT to understand the relation statements.

To learn the user-interested specific relation, we extract all the relation statements of user-given ⟨p0,c0⟩ as positive training samples. We collect negative training sentences in two ways:

- relation statements of sibling nodes, thus avoiding the model to only find closely related terms
- random sentences from the corpus, so the model can learn from irrelevant contexts to avoid overfitting.

Sequence Input Representation

Since user gives only a minimal number of seeds, we cannot simply add explicit markers around pairs of terms to let the model pay attention to.

Therefore, we take the original sentence and replace the two terms by "[MASK]" tokens with two justifications:

- aligning with the pre-trained objective of masked language model
- avoiding the classification layer to only remember relations from training pairs instead of looking into contexts.

Classification Layer

We take the output of two "[MASK]" tokens from the last layer of the pre-trained language model, and concatenate them to be the input of the classification layer, where we use a simple linear layer before the softmax layer.
The output label chooses the relation from e1 to e2 among three classes in the relation set Q: ⟨e1→e2⟩ (i.e., e1 is a parent node of e2), ⟨e2→e1⟩, or non-user interested relation.

Data Augmentation

To fully utilise the asymmetric property along the taxonomy edges, we augment each input training sequence by reversing the order of concatenation of e1 and e2. Then the label would switch if user-interested relation exists between the pair, but will not change otherwise.

First-layer Topic Finding by Root Node Discovery.

After deriving a relation classifier, we can easily transfer the user-interested relation along the paths in the taxonomy.
This is done by targeting an existing node and finding entities to be its potential parent node (transferring upwards) or child node (transferring downwards).
We assume that if we can discover potential root nodes, such as "Food" for "Dessert" and "Seafood", then the root node would have more general contexts for us to find connections with potential new topics.
We transfer the relation upwards by using the relation classifier learned to extract a list of parent nodes for each seed topic.
The common parent nodes for all topics are treated as root nodes R.

Finding common root nodes

To find the parent or child of an existing node, we extract relation statements of a concept e and a candidate term w into the relation classifier to judge sentence-based relations. Corpus-based relation between w and e is then averaged over confident sentence-based results over the corpus, with the confidence threshold being δ .

$$\text{Score}(w \rightarrow e) = \frac{\sum_{s_{w \rightarrow e}} \mathbb{1}\left(KL\left(\boldsymbol{l} \| \boldsymbol{p}_w\right) > \delta\right)}{\sum_{q \in Q} \sum_{s_q} \mathbb{1}\left(KL\left(\boldsymbol{l} \| \boldsymbol{p}_w\right) > \delta\right)}$$

Where

- sq denotes relation statements in which the relation of q exists
- pw denotes the output probability from the relation classifier
- l is the uniform distribution vector among three classes of relations.

Thus if the KL divergence between the two distributions is larger than a threshold δ , we

treat the prediction as a confident one.
The equation calculates the portion of term w being the parent of concept e among all the confident predictions, and we confirm w as the parent node of e if the portion is larger than a threshold.
For each user-given first-layer topic, we can generate a list of parent nodes, and their common parent nodes are treated as root nodes R.

Finding new first-layer topics
We apply the relation classifier to extract child terms for each root node r ∈ R.
This is done in a similar way as root node discovery, but we only reverse the direction of relation.
Thus we need to replace (w → e) in the previous equation with (r → w).
New topics are selected by their average score over all root nodes.

$$\text{Score}(R \rightarrow w) = \frac{\sum_{r \in R} \text{Score}(r \rightarrow w)}{|R|}$$

Candidate term extraction for subtopics
After generating the first-layer topics, we transfer the relation downwards to discover subtopics of each first-layer topic.
This can be done by applying the first equation again and replacing (w → e) with (e → w).
The candidate terms will later be clustered into subtopics.

Motivation

In this context, labels are generated to extended the (user provided) seed taxonomy.
Similarly to
`meng_2020_hierarchical_topic_mining_via_joint_spherical_tree_and_text_embedding`,
and contrary to most cases, seed and generated topic labels represent a starting point for topic modeling.

Topic modeling

Word embedding learning process guided by (extended) taxonomy.
(As part of the proposed seed-guided topical taxonomy construction)

Topic modeling parameters

- Embedding dimension: 100
- Local context window size: 5
- λd: 1.5
- λp: 1.0
- The threshold for Cluster Consistency: 0.5

Nr. of topics

Label

Single of multi word labels (taxonomy classes) taken from a pool of 16650 (DBLP) and 14619 (Yelp) terms.

Label selection

\

Label quality evaluation

\

Assessors

\

Domain

Paper: Topic taxonomies
Dataset: Scientific literature and Restaurants

Problem statement

Proposing a method for seed-guided topical taxonomy construction, taking:

- A corpus

- A seed taxonomy described by concept names

And:

- Constructing a more complete taxonomy
- Populating each node with a cluster of coherent terms.

Achieving this goal via:

- A relation transferring module (see label generation)
- A concept learning module that enriches the semantics for a taxonomy of words by extracting distinctive terms

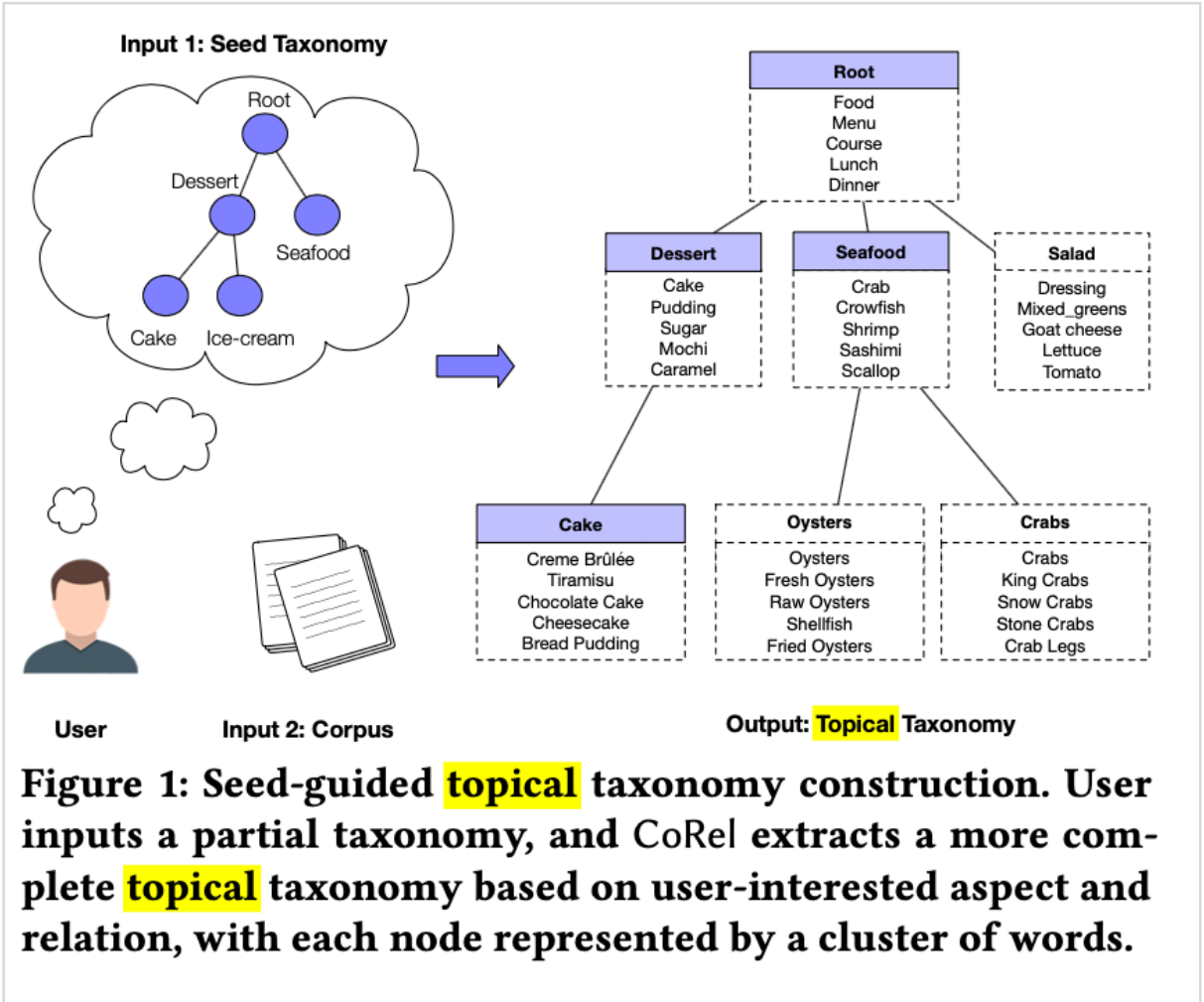

Figure 1: Seed-guided topical taxonomy construction. User inputs a partial taxonomy, and CoRel extracts a more complete topical taxonomy based on user-interested aspect and relation, with each node represented by a cluster of words.

Corpus

Dataset 1

Origin: DBLP

Nr. of documents: 157.000

Details: Abstracts from publications in the field of computer science

Dataset 2

Origin: Yelp

Nr. of documents: 1.08 million

Details: Restaurant reviews

Document

Dataset 1

Abstract related to a single paper.

Dataset 2

Single review of a restaurant

Pre-processing

- AutoPhrase to extract meaningful phrases to serve as our vocabulary.
- Infrequent terms occurring less than 50 times are discarded.

```
@inproceedings{huang_2020_corel_seed_guided_topical_taxonomy_construction_by_co
ncept_learning_and_relation_transferring,
author = {Huang, Jiaxin and Xie, Yiqing and Meng, Yu and Zhang, Yunyi and Han,
Jiawei},
title = {CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning
and Relation Transferring},
year = {2020},
isbn = {9781450379984},
publisher = {Association for Computing Machinery},
address = {New York, NY, USA},
url = {https://doi.org/10.1145/3394486.3403244},
doi = {10.1145/3394486.3403244},
abstract = {Taxonomy is not only a fundamental form of knowledge
representation, but also crucial to vast knowledge-rich applications, such as
question answering and web search. Most existing taxonomy construction methods
extract hypernym-hyponym entity pairs to organize a "universal" taxonomy.
However, these generic taxonomies cannot satisfy user's specific interest in
certain areas and relations. Moreover, the nature of instance taxonomy treats
```

each node as a single word, which has low semantic coverage for people to fully understand. In this paper, we propose a method for seed-guided topical taxonomy construction, which takes a corpus and a seed taxonomy described by concept names as input, and constructs a more complete taxonomy based on user's interest, wherein each node is represented by a cluster of coherent terms. Our framework, CoRel, has two modules to fulfill this goal. A relation transferring module learns and transfers the user's interested relation along multiple paths to expand the seed taxonomy structure in width and depth. A concept learning module enriches the semantics of each concept node by jointly embedding the taxonomy and text. Comprehensive experiments conducted on real-world datasets show that CoRel generates high-quality topical taxonomies and outperforms all the baselines significantly.},
booktitle = {Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining},
pages = {1928–1936},
numpages = {9},
keywords = {topic discovery, semantic computing, taxonomy construction, relation extraction},
location = {Virtual Event, CA, USA},
series = {KDD '20}
}

#Thesis/Papers/Initial