RaTE: a Reproducible automatic Taxonomy Evaluation by Filling the Gap

Tianjian Gao, Philippe Langlais
RALI/IDIRO, Université de Montréal
tianjian.gao@umontreal.ca,felipe@iro.umontreal.ca

Abstract

Taxonomies are an essential knowledge representation, yet most studies on automatic taxonomy construction (ATC) resort to manual evaluation to score proposed algorithms. We argue that automatic taxonomy evaluation (ATE) is just as important as taxonomy construction. We propose RaTE[1], an automatic label-free taxonomy scoring procedure, which relies on a large pre-trained language model. We apply our evaluation procedure to three state-of-the-art ATC algorithms with which we built seven taxonomies from the Yelp domain, and show that 1) RaTE correlates well with human judgments and 2) artificially degrading a taxonomy leads to decreasing RaTE score.

1 Introduction

A domain taxonomy is a tree-like structure that not only aids in knowledge organization but also serves an integral part of many knowledge-rich applications including web search, recommendation systems and decision making processes. Taxonomies are also inevitably used as business and product catalogs and for managing online sales. Notable taxonomy products in this domain include Amazon Category Taxonomy,[2] Google Product Taxonomy,[3] Yelp Business Category[4] and Google Content Categories.[5]

Recent years have witnessed interest in new automatic taxonomy construction (ATC) systems, but there are no systematic methods for objectively evaluating their figure of merit. For instance, TaxoGen (Zhang et al., 2018) — see Section 3 — was evaluated by asking at least three human evaluators if a taxonomy concept pair contains a hypernymy relationship, which can lead to bias and low reproducibility. It is not only difficult to compare or rank different algorithms, but changing the hyperparameters or settings of a parameterized ATC system can also result in drastically different outputs, and make optimization unfeasible.

Because ontologies and taxonomies in particular are typically created in contexts to address specific problems or achieve specific goals, e.g. classification, their evaluation is evidently context-dependent, and many researchers actually believe that a task-independent automatic evaluation remains elusive (Porzel and Malaka, 2004). Still, researchers have argued that objective evaluation metrics must be well available for significant progress in the development and deployment of taxonomies and ontologies (Brewster et al., 2004).

In this work, we propose RaTE, a Reproducible procedure for Automatic Taxonomy Evaluation. RaTE does not require external knowledge but instead depends on masked language modelling (MLM) to query a large language model for subsumption relations. We show that with some care, MLM is a valuable proxy to human judgments.

We apply RaTE to the Yelp corpus (a corpus of restaurant reviews) ranking seven taxonomies we extracted using three state-of-the-art ATC systems. We observe it correlates well with our manual evaluation of those taxonomies, and also show that artificially degrading a taxonomy leads to a decrease of score proportional to the level of noise injected.

In the remainder, we discuss related work in Section 2. In Section 3, we describe the ATC systems we used for building up our taxonomies, and their evaluation procedures. We then present RaTE in Section 4 including refinements that we found

[1] Our code repository is available at https://github.com/CestLucas/RaTE

[2] https://www.data4amazon.com/amazon-product-taxonomy-development-mapping-services.html

[3] https://support.google.com/merchants/answer/6324436?hl=en

[4] https://blog.yelp.com/businesses/yelp_category_list/

[5] https://cloud.google.com/natural-language/docs/categories?hl=fr

necessary for our approach to work. We report in Section 5 the experiments we conducted to demonstrate the relevance of RaTE, and conclude in Section 6.

2 Related Works

Systematic methods of evaluating ontologies and taxonomies are lacking. Because agreed upon quantitative metrics are lacking, research on taxonomy and ontology construction relies heavily on qualitative descriptions and the various perspectives of ontology engineers, system users or domain experts, which renders the results subjective and unreproducible (Gómez-Pérez, 1999; Guarino, 1998).

Brank et al. (2005) summarized four principle ontology evaluation methods, by (1) comparing the target ontology to a "gold standard" (ground-truth) ontology (Maedche and Staab, 2002); (2) using the target ontology in an application and evaluating the application results ("application based") (Porzel and Malaka, 2004); (3) conducting coverage analysis comparing the target with a source of data (eg., a collection of documents) about a specific domain ("data driven") (Brewster et al., 2004); (4) manual reviews done by human experts that assess how well the target ontology meets a set of predefined criteria, standards, and requirements (Lozano-Tello and Gómez-Pérez, 2004).

Gold Standard Evaluation focusses on comparing and measuring the similarity of the target taxonomy with an existing ground truth such as WordNet (Fellbaum, 1998), Wikidata and ResearchCyc (Ponzetto and Strube, 2011). Semantic similarity metrics have been proposed, including Wu-Palmer (Wu and Palmer, 1994), Leacock-Chodorow (Leacock and Chodorow, 1998) and Lin (Lin, 1998). We include in this category specific measures such as *topic coherence* (Newman et al., 2010) which scores the quality of a word cluster which rely on similarity measures. There are several issues with such a process: mapping concepts from the output system to the ground truth is not trivial and gold standards do not necessarily cover well the domains of interest.

Application-based Evaluation is an attractive alternative to gold-standard evaluation. Porzel and Malaka (2004) for instance proposed several possible applications for evaluation including concept-pair relation classification. Brank et al. (2005) underlines however that it is in fact hard to correlate ontology quality with the application performance.

Data-driven Evaluation intends to select the ontology O with the best structural *fit* to a target corpus C, which boils down into estimating $P(C|O)$ as in (Brewster et al., 2004). Practically however, it remains unclear how to approximate such conditional probability.

3 Automatic Taxonomy Extractors

In this work, we replicated results of three state-of-the-art ATC systems that are publicly available and that are producing quality results on selected datasets and domains. In this section, we describe those systems and discuss their corresponding evaluation methods.

3.1 TaxoGen

TaxoGen (Zhang et al., 2018) is an adaptive text embedding and clustering algorithm leveraging various phrase-mining and clustering techniques including AutoPhrase (Shang et al., 2018), caseOLAP (Liem et al., 2018) and spherical k-means clustering (Banerjee et al., 2005). TaxoGen iteratively refines selected keywords and chooses cluster representative terms based on two criteria: *popularity* which prefers term-frequency in a cluster and *concentration* which assumes that representative terms should be more relevant to their belonging clusters than their sibling clusters.

The system can be configured with several hyperparameters including the depth of the taxonomy, the number of children per parent term and the "representativeness" threshold. Experiments were conducted on DBLP and SP (Signal Processing) datasets and the system is quantitatively evaluated with relation accuracy and term coherency measures assessed by human evaluators (10 doctoral students).

3.2 CoRel

CoRel (Huang et al., 2020) takes advantages of novel relation transferring and concept learning techniques and uses hypernym-hyponym pairs provided in a seeded taxonomy to train a BERT (Devlin et al., 2019) relation classifier and expand the seeded taxonomy horizontally (width expansion) and vertically (depth expansion). Topical clusters are generated using pre-computed BERT embeddings and a discriminative embedding space is learned, so that each concept is surrounded by its representative terms.

The clustering algorithms used by CoRel are *spectral co-clustering* (Kluger et al., 2003) and *affinity propagation* (Frey and Dueck, 2007), which automatically computes the optimal number of topic clusters. Compared to TaxoGen, CoRel does not require depth and cluster number specifications but a small seeding taxonomy as an input for enabling a weakly-supervised relation classifier.

CoRel is quantitatively evaluated with term coherency, relation F1 and sibling distinctiveness judged by 5 computer science students on subsets of DBLP and Yelp datasets. The system generates outputs in the form of large hierarchical topic word clusters.

3.3 HiExpan

HiExpan (Shen et al., 2018) is a hierarchical tree expansion framework that aims to dynamically expand a seeded taxonomy horizontally (width expansion) and vertically (depth expansion) and performs entity linking with Microsoft's Probase (Wu et al., 2012) — a probabilistic framework used to harness 2.7 million concepts mined from 1.68 billion web pages — to iteratively grow a seeded taxonomy. As an entity is matched against a verified knowledge base, we perceive the accuracy of terms and concept relations to be higher than that of CoRel and TaxoGen.

Authors of the HiExpan, as well as some volunteers assessed the taxonomy parent-child pair relations using ancestor- and edge-F1 scores.

3.4 Observations

Each of those taxonomy extractors face their own set of advantages and drawbacks. TaxoGen is the only parameterized systems in our experiments, and is the only one that does not require a seeded input for producing an output, which can be beneficial when prior knowledge of the corpus is lacking. It also generates alternative synonyms for each taxonomy topic, which increases the coverage and improves concept mapping between taxonomies and documents. However, it seems to depend on the keyword extraction quality and it is unclear how to determine the best hyper-parameter settings owing to the lack of automatic evaluation methods.

CoRel uses the concept pairs provided in the seed taxonomy for mining similar relations, but this has become its Achilles' heel because same-sentence co-occurrence of valid parent-child topics is rare in real-world data. As a result, CoRel may fail to produce any output at all due to insufficient training examples for the relation classifier. It is also resource-intensive for making use of neural networks for relation transferring and depth expansion. Anecdotally, the output of CoRel may also not be entirely exhaustive and deterministic.

For our experiments, HiExpan is perceived to produce the most consistent taxonomies thanks to the use of Probase for measuring topic similarities and locating related concepts. However, the set-expansion mechanism of HiExpan often ignores topic granularity and adds hyponyms and hypernyms found in similar contexts to the exact same taxonomy level (hence most HiExpan taxonomies are two-level only). It also cannot differentiate word senses such as virus as in *computer virus* and a *viral disease*.

4 RaTE

A critical part of taxonomy/ontology evaluation is knowledge about subsumptions, e.g. "is *fluorescence spectroscopy* a type of *fluorescence technology*?" or "is *CRJ200* a *Bombardier*?".

Thus, RaTE measures the accuracy of the hypernym relations present in a taxonomy we seek to evaluate. The main difference between our work and earlier ones is that we do not rely on human judgments to determine the quality of a parent-child pair, nor do we consider an external reference (that often is not available or simply too shallow). Instead, we rely on a large language model tasked to check subsumption relations.

Ultimately, an optimized language model should be able to generate an accurate list of the most canonical hypernyms for a given domain, similar to domain experts. But because we are mainly interested in domain-specific taxonomies, there is a high risk that specific terms of the domain are not well recognized by the model, and therefore, we investigate three methods for increasing the hit rate of hypernymy prediction of taxonomy subjects and reducing false negatives by (1) creating various prompts, (2) fine-tuning MLMs with different masking procedures, and (3) extending the model's vocabulary with concept names.

4.1 Core idea

We consider a taxonomy as a set of n parent-child pairs from adjacent taxonomy levels linked by single edges, denoted as $(p, c) \in \mathcal{T}$. For each parent-child pair $(p_i, c_i), i \in 1, ..., n$, we insert c_i and the "[MASK]" token into some prompts containing

c	Pred 1	Pred 2	Pred 3	Pred 4	Pred 5	Rank
Mussel	fish (0.227)	dish (0.144)	seafood (0.140)	meat (0.037)	soup (0.033)	3
Clam	fish (0.203)	dish (0.095)	seafood (0.076)	crab (0.030)	thing (0.027)	3
Lobster	seafood (0.222)	dish (0.145)	lobster (0.131)	food (0.052)	sauce (0.052)	1
Chicken	dish (0.167)	meat (0.110)	chicken (0.079)	thing (0.058)	sauce (0.052)	73
Beef	meat (0.274)	beef (0.161)	dish (0.063)	food (0.027)	thing (0.024)	57

Table 1: Top-5 hypernym predictions made by a pre-trained BERT model (Bert-large-uncased-whole-word-masking) by prompting it with "*c is a type of [MASK]*". The rank of seafood in the list is indicated in the last column.

"is-a" patterns (Hearst, 1992), then use LMs to unmask $p'_1(c_i), p'_2(c_i), ..., p'_k(c_i) \in p'(c_i)$ per query as proxy parent terms of c_i, where k is a recall threshold (we used $k = 10$ in this work). This process is illustrated in Table 1.

A good pair of taxonomy concepts is therefore if the parent concept p_i can be found among the machine predictions $p'(c_i)$. We consider a parent-child relation *positive* if and only if the parent term is recalled one or more[6] times in the top k predictions. This policy can obviously be adjusted, which we leave as future work. The measure of quality of $\mathcal{T}$ is then simply the percentage of (p, c) links in $\mathcal{T}$ that are correct according this procedure. We note that for a taxonomy with no parent-child pairs, i.e. a single-level taxonomy, our evaluation score is 0.

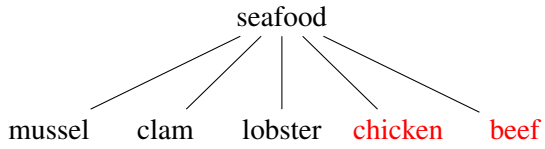

Figure 1: Excerpt from HiExpan1 for topic "seafood"

As an illustration, the taxonomy in Figure 1 would receive a score of 3/5 based on the predictions made in Table 1 where for instance, $p'_1(c_i), p'_2(c_i), ..., p'_5(c_i)$ equal *fish, dish, seafood, meat, soup* for c_i = *mussel*, in which we find the real taxonomy parent p_i = *seafood* = $p'_3(c_i)$.

We observe from Table 1 that not every prediction is factually correct (e.g. mussels are neither fish nor meat), and it remains evidently unreliable to depend solely upon pre-trained language models as ground-truth for all knowledge domains. Yet, we argue that we can regard the rankings of MLM predictions as a likelihood of a subsumption relation between the subject and the object of a query. In our example, the model is significantly more likely to predict "seafood" for *mussel, clam* and *lobster* (rank 3,3,1) than for *chicken* and *beef* (rank 73,57).

4.2 Diversified Prompting

Models can produce all sorts of trivial predictions, such as stop-words (e.g. "**this** is a kind of seafood"), or expressions and collocations found frequently in training samples (e.g. "seafood is a kind of **joke/disappointment**").

Differences in prompts used can actively impact a model's performance in hypernymy retrieval (Peng et al., 2022; Hanna and Mareček, 2021). Hanna and Mareček (2021) reported that prompting BERT for hypernyms can actually outperform other unsupervised methods even in an unconstrained scenario, but the effectiveness of it depends on the actual queries. For example, they show that the query "A(n) x **is a** [MASK]" outperformed "A(n) x **is a type of** [MASK]" on the Battig dataset.

As a result, instead of relying on a single query, we design five pattern groups (p1-p5) of hypernymy tests for pooling unmasking results. Those are illustrated in Table 2 for the parent-child pair (seafood,shrimp).

While p2 to p4 follow standard Hearst-like patterns (Hearst, 1992), p5a employs the "my favourite is" prompt which has demonstrated high P@1 and MRR in (Hanna and Mareček, 2021). Patterns p1 have been created specifically for noun phrases that have a tendency to be split and considered as good taxonomy edges by ATC systems.[7]

With this refined set of patterns, a topic pair has therefore a score of 1, as in the seafood-shrimp example, if the parent term is among the top-k machine predictions for any inquiries containing the child topic, and 0 vice versa. Again, more elaborate decisions can be implemented.

[6] A parent word can be predicted multiple times in singular and plural forms, misspellings, and so on, e.g. "dessert", "desserts" and "desert".

[7] For instance, extractors tend to produce (salad,shrimp) for the pair (salad,shrimp salad).

Prompt		Pred1	Pred2	Pred3	Pred4	Pred5	Rank
p1a	{shrimp} [MASK]	salad	cocktail	pasta	soup	rice	359
p1b	[MASK] {shrimp}	fried	no	garlic	coconut	fresh	117
p2a	{shrimp} is a [MASK]	joke	must	winner	favorite	hit	959
p2b	{shrimp} is an [MASK]	option	issue	experience	art	order	4407
p3a	{shrimp} is a kind of [MASK]	joke	thing	dish	treat	disappointment	146
p3b	{shrimp} is a type of [MASK]	dish	thing	food	sauce	seafood	5
p3c	{shrimp} is an example of [MASK]	that	this	shrimp	food	seafood	5
p4a	[MASK] such as {shrimp}	sides	food	seafood	fish	shrimp	3
p4b	A [MASK] such as {shrimp}	lot	variety	side	combination	protein	40
p4c	An [MASK] such as {shrimp}	ingredient	item	option	order	animal	197
p5a	My favorite [MASK] is {shrimp}	dish	thing	part	item	roll	16

Table 2: Evaluation queries for the parent-child pair (seafood,shrimp).

4.3 Fine-tuning the Language Model

To improve hypernymy predictions, we must also address two issues with pre-trained language models: (1) the models are untrained on the evaluation domain; (2) the default model tokenizer and vocabulary are oblivious of some taxonomy topics, resulting in lower recall.

Most research on MLM prompting only assessed the performance of pre-trained models. Yet, Peng et al. (2022) found an improvement when using FinBert models (Yang et al., 2020) pre-trained with massive financial corpora in retrieving financial hypernyms such as *equity* and *credit* for *"S&P 100 index is a/an __ index"*, compared to using BERT-base. Also, Dai et al. (2021) generated ultra-fine entity typing labels, e.g. "person, soldier, man, criminal" for *"**he** was confined at Dunkirk, escaped, set sail for India"* through inserting hypernym extraction patterns and training LMs to predict such patterns.

Analogously, we compared six fine-tuned models, investigating different masking protocols, model vocabulary (see next section) and training sizes. Because we want the language models to concentrate on the taxonomy entities, particularly the parent terms and their surrounding contexts, we prioritize therefore masking the main topics (shown in Table 3) and parent terms of the taxonomies to evaluate, then other taxonomy entities (e.g. leaf nodes), followed by AutoPhrase entities if no taxonomy entities are present in the sentence and other random tokens from our training samples. In addition, we test entity masking by only masking *one* taxonomy entity rather than 15% of sentence tokens to gain more sentence contexts. Our masking procedures are illustrated in Figure 2.

4.4 Extended Vocabulary

Domain-specific words such as food items are typically not predicted as a whole word, but rather as a sequence of subword units, such as *appetizer* which is treated as *'app', '##eti' and '##zer'* by the standard tokenizer. To avoid multi-unit words to be overlooked by the language model, we propose to extend its vocabulary.

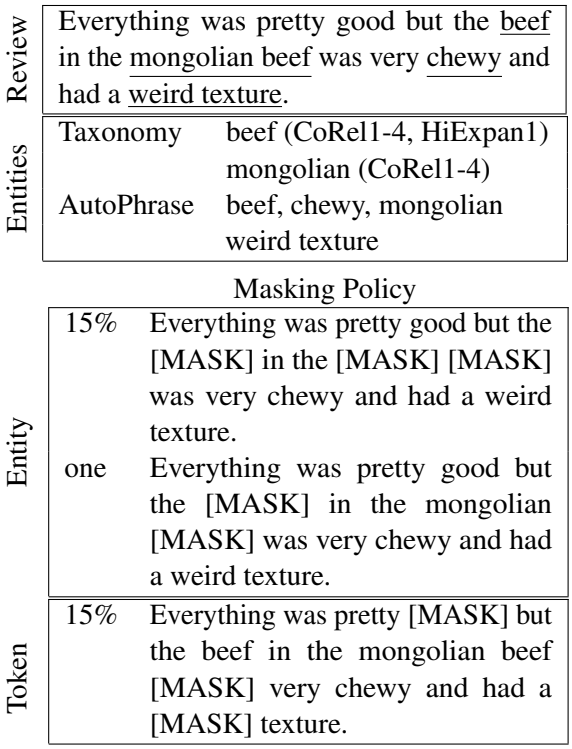

Figure 2: Comparison of masking strategies for a sample Yelp review where taxonomy entities or those proposed by AutoPhrase are underlined. We prioritize masking the taxonomy entities, AutoPhrase entities and random tokens, in that order.

We enrich the vocabulary of models m1 and m2, by adding the lemmas (or singular forms) of parent terms from Table 3 that were not previously

included in the base tokenizer, such as “sushi", “appetizer" and “carne asada", and resizing the models' token embedding matrices to match the size of the new tokenizer. The embedding representation of new tokens were initialized randomly before fine-tuning, although it is possible to assign them to the representation of the closest terms in the original vocabulary.

By adding only a small number of new tokens to the model and tokenizer, we also ensure similar model and tokenizer efficiencies. We believe that vocabulary extension will become a necessary step for effective hypernymy prediction in most specialized domains, though the exact optimal strategies remain to be discussed.

5 Experiments

We conducted our experiments on the Yelp corpus which contains around 1.08M restaurant reviews such as the one in Figure 2 (top box). We used the very same corpus prepared by Huang et al. (2020).[8]

5.1 Taxonomies

We created seven taxonomies for evaluation using the ATC systems mentioned in Section 3. Here our goal was to obtain meaningful taxonomies that best cover the Yelp domain using each taxonomy extractor. We did so by experimenting with different extractor settings and input. For TaxoGen, we only had to specify some parameters.[9] For CoRel and HiExpan however, we had to provide a seed taxonomy. Hence we produced 5 such taxonomies using CoRel and HiExpan by providing frequently-appearing parent-child pairs in the seeds.[10]

Table 3 reports the main topics (level 1) of the produced taxonomies. We observe that the output of one ATC system varies substantially from one parametrization to another. Also, it is noticeable that the main topic of some taxonomies do lack structure. For instance, putting *beef* and *chicken* in the category *meat* would arguably make better sense in CoRel1.

5.2 Models

We fined-tuned six language models according to the different strategies we presented in Section 4 and which characteristics are summarized in Table 4. In particular, we experiment with *entity masking* while fine-tuning model m1a, m1b and m0b, which emphasizes masking task-relevant tokens, because it has been shown to be more effective than *random masking* in (Sun et al., 2019; Kawintiranon and Singh, 2021). All models have been fine-tuned for 2 epochs by masking 15% tokens, to the exception of m1b (marked with $\star$) for which only one entity has been masked per example.

Taxonomy	Top level (main) topics
CoRel1	steak, veggies, beef, cheese, crispy, fish, rice, salad, shrimp, spicy, pork, bacon, burger, appetizer, bread, dessert, seafood
CoRel2	bacon, bread, fries, roll, soup, burger, dessert, salad, shrimp
CoRel3	chinese, seafood, dessert, steak
CoRel4	dinner, food, location, lunch, service
HiExpan1	seafood, salad, dessert, appetizer, food, sushi, desert, pizza, coffee, bread, pasta, beer, soup, wine, cheese, cocktail, taco, water, music
TaxoGen1	main_dish, south_hills, high_ceilings, était_pas
TaxoGen2	chest, tempe, amaretto, pepper_jelly, relies, travis, free_admission, exposed_brick

Table 3: Main targets of MLM evaluation.

Model name (base)	Finetuning Voc.	Full	70%	Masking Ent.	Tok
m1a (bert-base)	✓	✓		✓	
m1b (bert-base)	✓	✓		✓$\star$	
m2a (bert-base)	✓	✓			✓
m2b (bert-base)	✓		✓		✓
m0a (bert-base)			✓		✓
m0b (distilbert-base)			✓	✓	

Table 4: Configurations of the fine-tuned models, with models m0a and m0b serving as baselines for training with the base tokenizer; m0b using a smaller pre-trained model and less fine-tuning material. Column Voc. indicates that main target words proposed ATC systems were injected in the model's vocabulary.

For comparison purposes, we also selected two pre-trained models bert-large-uncased-whole-word-masking and bert-base-uncased that we did not fine-

[8] Available at: https://drive.google.com/drive/folders/13DQ0II9QFLDhDbbRcbQ-Ty9hcJETbHt9.

[9] We considered taxonomy depth, number of topics per level, and “word filtering threshold”. See the github for the specific values we used.

[10] They pretty much align with the one used by Huang et al. (2020), although we proceeded by trial-error until satisfaction.

Model	Pred1	Pred2	Pred3	Pred4	Rank
m1a	burger	dish	sandwich	steak	4
m1b	dish	burger	beer	sandwich	10
m2a	steak	dish	meat	cut	1
m2b	steak	dish	burger	meat	1
m0a	dish	burger	steak	meat	3
m0b	cut	steak	meat	beef	2
B-l	fruit	flavor	food	color	69
B-b	food	drink	color	dessert	71

Table 5: Fine-tuned (top) vs. pre-trained (bottom) models' top-4 predictions with the prompt "my favourite [MASK] is sirloin ."

Model	Pred1	Pred2	Pred3	Pred4	Rank
m1a	sides	foods	food	apps	5
m1b	sides	food	appetizer	foods	3
m2a	sides	items	food	dessert	6089
m2b	things	items	foods	props	3111
m0a	sides	extras	items	dessert	N/A
m0e	sides	apps	foods	snacks	N/A
B-l	foods	items	products	food	N/A
B-b	foods	snacks	food	items	N/A

Table 6: Top-4 predictions of models with extended (top) or base (bottom) vocabulary for the prompt "[MASK] such as mozzarella sticks".

tune and that we named B-l and B-b respectively.

To highlight the qualitative differences between our evaluation models, we provide a simple prompt "my favourite [MASK] is sirloin" for the models to predict the taxonomy hypernym "steak" in CoRel1. The results are shown in Table 5, where 5 out of 6 fine-tuned models and none of the pre-trained models correctly predicted the taxonomy parent in the top 4 predictions. Further, all fine-tuned models returned "steak" in the top ten predictions.

Lastly, we show the positive effects of extending the vocabulary of the language model in Table 6 where we wish to recall the parent term "appetizer" for the concept pair "appetizer-mozzarella sticks" in CoRel1, where the token "appetizer" would be split into *'app', '##eti' and '##zer'* by the standard tokenizer. Both models m1a and m1b trained with entity masking and an expanded vocabulary correctly predicted "appetizer" in their top five predictions; m2 models also recalled the term, albeit with a very low rank whereas other models are completely oblivious to it. Nevertheless, we find that expanding the model's vocabulary in conjunction with entity masking may introduce bias into the models when fine-tuning with limited training samples, i.e. over predicting the added tokens.

5.3 Ranking Results

5.3.1 Manual Ranking

The first author of this paper first manually ranked the extracted taxonomies prior to experimenting with RaTE. The main task was to manually verify the validity of the parent-child pairs of each taxonomy, while also taking into account factors like taxonomy structure.[11]

HiExpan1 was deemed the best taxonomy, likely because the word relations actually originate from a verified database and the coverage is extensive. It is also observably more accurate than CoRel 1-4, which have similar (overall good) quality. TaxoGen taxonomies were the least accurate, with TaxoGen1 superior to TaxoGen2. We found them trivial in the sense that the algorithm selects many insignificant topics because no seeded taxonomy indicating user interest is provided. We believe that another cause for this is the system's low sensitivity to keywords supplied by AutoPhrase, which on Yelp generates too many irrelevant terms and leads to many noisy concept pairs (e.g. "exposed brick – music video").

In fact, manually ranking the HiExpan and TaxoGen taxonomies was simple and obvious, but ranking the CoRel taxonomies was more complex. Such an assessment is delicate; after all, this was the principal motivation of RaTE.

5.3.2 RaTE Ranking

Table 7 showcases the results of MLM taxonomy relation accuracy evaluation, calculated by the number of positive relations over all unique parent-child pairs in a taxonomy.[12]

The entity-masking models m1a and m1b predicted the most positive relationships in each candidate taxonomy while the pre-trained models predicted the fewest, which was expected. It is also surprising that B-b outperforms B-l when it comes to matching more positive concept pairs. Model m2b (trained on two-thirds of the data) expectedly

[11] All HiExpan1 and TaxoGen1&2 parent-child pairs were manually examined, however due to the large size of the word clusters, we had to sample and evaluate concept pairs for CoRel 1-4.

[12] We considered word inflections and certain special cases to improve matching between taxonomy terms and machine predictions, e.g. "veggies", "vegetable" and "vegetables"; "dessert" and "desert".

	Fine-tuned Models						BERT		Majority	RaTE	Manual
	m1a	m1b	m2a	m2b	m0a	m0b	large	base	Voting	ranking	ranking
CoRel1	72.7	71.8	42.4	44.5	46.3	43.6	20.4	27.4	44.3	4	3
CoRel2	78.2	75.0	54.4	53.7	**57.2**	51.2	25.9	36.2	57.2	2	2
CoRel3	60.2	66.7	54.1	54.9	**57.2**	50.1	36.0	40.0	53.5	3	4
CoRel4	68.2	64.6	45.0	39.0	36.5	38.1	**41.0**	41.8	34.7	5	5
HiExpan1	**84.5**	**84.7**	**59.5**	**56.7**	56.9	**64.3**	34.9	**42.0**	**59.0**	1	1
TaxoGen1	13.5	14.7	5.5	6.1	1.2	2.5	3.1	3.7	1.2	6	6
TaxoGen2	0.0	0.0	0.0	5.0	0.0	0.0	0.0	0.0	0.0	7	7

Table 7: Relation accuracy scores evaluated by language models, calculated by the number of positive relations, or parent terms in the model predictions, divided by the number of unique parent-child pairs in each taxonomy.

underperforms model m2a, but not drastically.

However, all models produce overall similar score distributions, with the HiExpan taxonomy receiving the highest scores and the TaxoGen taxonomies receiving the lowest. This is consistent with our manual judgements in that the HiExpan concept pairs were derived from an accurate relation dataset (Probase), whereas TaxoGen1 and TaxoGen2 contain mostly noise.

We also compute the majority voting scores for each evaluation target using the six models of Table 4: a concept pair of a taxonomy is positive if and only if three or more models have successfully predicted the parent word. The resulting ranking is reported in the next column, and is shown to correlate well with our manual evaluation (last column).

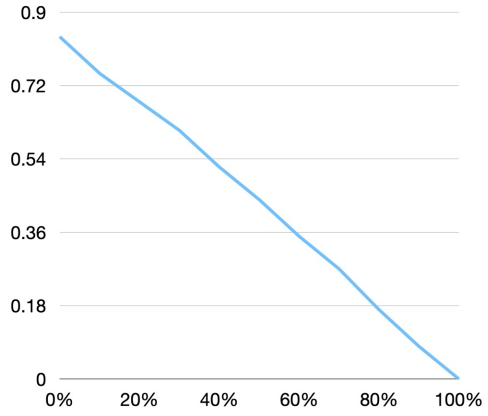

Figure 3: Relation accuracy obtained with model m1a, as a function of the percentage of noise introduced in HiExpan1.

5.4 Random noise Simulation

To further evaluate the good behaviour of RaTE, we conducted an experiment where we degraded the HiExpan1 taxonomy (the best one we tested). We did this by randomly replacing a percentage of concepts by others. Figure 3 shows that the score (obtained with model m1a) roughly decreases linearly with the level of noise introduced, which is reassuring.

6 Discussion

We presented RaTE, a procedure aimed at automatically evaluating a domain taxonomy without gold standard references or human evaluations. It relies on a large language model and an unmasking procedure for producing annotations. We tested RaTE on the Yelp corpus which gathers restaurant reviews, and found that it correlated well with human judgments, and (artificially) degrading a taxonomy led to a score degradation proportional to the amount of noise injected. Still, we observed that the quality of the language model predictions varies according to the strategies used to fine-tune them.

There remains a number of avenues to investigate. First, we have already identified a number of decisions that could be revisited. In particular, we must test RaTE on other domains, possibly controlling variables such as the size of the fine-tuning material or the frequency of terms. Second, RaTE is an accuracy measure, and depending on the evaluation scenario, it should eventually be coupled with a measure of recall. Last, an interesting avenue is to investigate whether RaTE can be used to optimize the hyper-parameters of an ATC system.

Acknowledgments

Our work has been done in collaboration with IATA, to whom we are truly grateful. We would like to thanks Olena Vasylchenko, Hyuntae Jung and Sorina Radu in particular for their support.

References

2011. Taxonomy induction based on a collaboratively built knowledge repository. *Artif. Intell.*, 175(9-10):1737–1756.

Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. 2005. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *Journal of Machine Learning Research*, 6(9).

Janez Brank, Marko Grobelnik, and Dunja Mladenić. 2005. A Survey of Ontology Evaluation Techniques. In *Proc. of 8th Int. multi-conf. Information Society*, pages 166–169.

C. Brewster, H. Alani, S. Dasmahapatra, and Y. Wilks. 2004. Data Driven Ontology Evaluation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 641–644, Lisbon, Portugal.

Hongliang Dai, Yangqiu Song, and Haixun Wang. 2021. Ultra-Fine Entity Typing with Weak Supervision from a Masked Language Model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1790–1799, Online.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota.

Fellbaum. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. The MIT Press.

Brendan J Frey and Delbert Dueck. 2007. Clustering by passing messages between data points. *Science*, 315(5814):972–976.

Asunción Gómez-Pérez. 1999. Evaluation of taxonomic knowledge in ontologies and knowledge bases. In *Banff Knowledge Acquisition for Knowledge-Based Systems (KAW'99)*, pages 6.1.1–6.1.18.

Nicola Guarino. 1998. Some ontological principles for designing upper level lexical resources. *arXiv preprint cmp-lg/9809002*.

Michael Hanna and David Mareček. 2021. Analyzing BERT's Knowledge of Hypernymy via Prompting. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 275–282.

Marti A Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Jiaxin Huang, Yiqing Xie, Yu Meng, Yunyi Zhang, and Jiawei Han. 2020. CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1928–1936.

Kornraphop Kawintiranon and Lisa Singh. 2021. Knowledge Enhanced Masked Language Model for Stance Detection. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 4725–4735.

Yuval Kluger, Ronen Basri, Joseph T Chang, and Mark Gerstein. 2003. Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions. *Genome research*, 13(4):703–716.

Claudia Leacock and Martin Chodorow. 1998. Combining Local Context and WordNet Similarity for Word Sense Identification. *WordNet: An electronic lexical database*, 49(2):265–283.

David A Liem, Sanjana Murali, Dibakar Sigdel, Yu Shi, Xuan Wang, Jiaming Shen, Howard Choi, John H Caufield, Wei Wang, Peipei Ping, et al. 2018. Phrase mining of textual data to analyze extracellular matrix protein patterns across cardiovascular disease. *American Journal of Physiology-Heart and Circulatory Physiology*, 315(4):H910–H924.

Dekang Lin. 1998. An Information-Theoretic Definition of Similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 98)*, pages 296—-304, San Francisco, CA, USA.

Adolfo Lozano-Tello and Asunción Gómez-Pérez. 2004. ONTOMETRIC: A method to choose the appropriate ontology. *Journal of Database Management (JDM)*, 15(2):1–18.

Alexander Maedche and Steffen Staab. 2002. Measuring similarity between ontologies. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 251–263. Springer.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic Evaluation of Topic Coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pages 100–108.

Bo Peng, Emmanuele Chersoni, Yu-Yin Hsu, and Chu-Ren Huang. 2022. Discovering Financial Hypernyms by Prompting Masked Language Models. In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pages 10–16.

Robert Porzel and Rainer Malaka. 2004. A Task-based Approach for Ontology Evaluation. In *ECAI Workshop on Ontology Learning and Population, Valencia, Spain*, pages 1–6. Citeseer.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated Phrase Mining from Massive Text Corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1825–1837.

Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle T. Vanni, Brian M. Sadler, and Jiawei Han. 2018. HiExpan: Task-Guided Taxonomy Construction by Hierarchical Tree Expansion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2180–2189. Association for Computing Machinery.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. ERNIE: Enhanced Representation through Knowledge Integration. *arXiv preprint arXiv:1904.09223*.

Wentao Wu, Hongsong Li, Haixun Wang, and Kenny Q Zhu. 2012. Probase: A Probabilistic Taxonomy for Text Understanding. In *Proceedings of the 2012 ACM SIGMOD international conference on management of data*, pages 481–492.

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.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. FinBERT: A Pretrained Language Model for Financial Communications. *arXiv preprint arXiv:2006.08097*.

Chao Zhang, Fangbo Tao, Xiusi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. Taxogen: Unsupervised Topic Taxonomy Construction by Adaptive Term Embedding and Clustering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2701–2709.