Unsupervised Common Sense Relation Extraction

Anonymous ACL submission

Abstract

Vast and diverse knowledge about the relations in the world help humans comprehend and argue about their environment. Equipping machines with this knowledge is challenging yet essential for general reasoning capabilities. Here, we propose to apply unsupervised relation extraction (URE), aiming to induce general relations between concepts from natural language. Previous work in URE has predominantly focused on relations between named entities in the encyclopedic domain. The more general, and more challenging, domain of *common sense* relation learning has not yet been addressed, partially due to a lack of datasets. We present a framework for common sense relation extraction from free-text, associated with two benchmark datasets. We present initial experiments using three state-of-the-art models developed for encyclopedic relation induction. Our results verify the utility of our benchmarks for common sense relation extraction, and suggest ample scope for future work on this important, yet challenging, task.[1]

1 Introduction

Humans possess a vast repository of basic facts and relations, which they use to perceive, navigate, reason about their environment – a resource called common sense knowledge. For instance, humans know that '*eating* is the FUNCTION of *forks*', or '*being scared* is the EMOTIONAL EVALUATION of seeing a *ghost*'.[2] Equipping machines with similar resources has attracted substantial attention in recent years (Davis and Marcus, 2015), for instance by incorporating existing resources (like `ConceptNet`; Liu and Singh (2004)) into models to solve downstream tasks like question answering (Lin et al., 2019); or by leveraging large

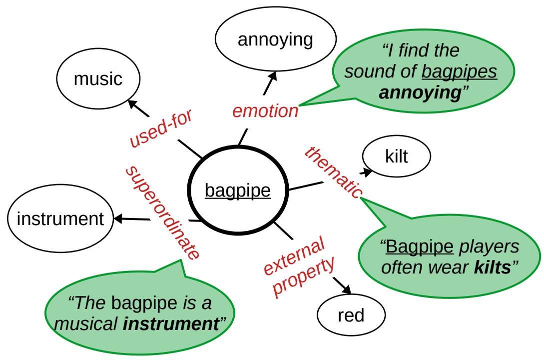

Figure 1: Illustration of WAREL, which consists of associations between cue words (bagpipe) and associations (kilt, red, ...) together with association explanations (speech bubbles) and discrete relation type labels (arrow labels).

pre-trained language models as common sense resources (Davison et al., 2019; Petroni et al., 2019; Shwartz et al., 2020). Prior work predominantly focussed on the fact *that* concepts are related, but less so on the specific *relations between* concepts. However, scalable knowledge of common sense relations is likely to benefit common sense reasoning applications. This paper introduces the task of common sense relation extraction.

Given the broad nature of common sense knowledge, manual collection of exhaustive concept relation data bases is infeasible. Instead, we follow recent work in the encyclopedic domain (Yao et al., 2011; Marcheggiani and Titov, 2016; Tran et al., 2020), and infer common sense relations between pairs of concept from concept mentions in text. Intuitively, given a corpus of sentences which mention pairs of concepts, we want to learn a small number of underlying common sense relations which explain the associations between the two concepts. Examples of common relations include USED-FOR, MADE-OF, or LOCATION, and relation inventories used in this work are discussed further in § 3. In the encyclopedic domain, rele-

[1]Code and data will be made publicly available upon acceptance under a CC BY SA 4.0 license.

[2]We denote *concepts* in italics, and RELATIONS in small caps throughout the paper.

vant corpora have been constructed using templates and heuristic supervision (Yao et al., 2011), however, the quality of the resulting data sets has been shown to be low (Gao et al., 2021). This problem is exacerbated in the common sense scenario where relations are broader, and while encyclopedic relations typically concern named entities, common sense relations span concepts, actions, properties and more. The core contribution of this paper are two sizeable, English data sets with complementary strengths to train and test common sense relation extraction models.

First, CNREL (Table 1, top) is based on `ConceptNet` (Speer et al., 2017), where we associate relation-labelled concept pairs with natural language sentences from the OMCS data set (Singh et al., 2002) using heuristic supervision. This data set is large, yet potentially noisy as sentences are not guaranteed to express the intended relation. In addition OMCS sentences are often templated.

Second, we collected a novel data set, WAREL (Table 1, bottom), which encodes relational human common sense knowledge through word associations (Deyne et al., 2019; Liu et al., 2021a). In a large crowd-sourcing study, we (a) collected human concept associations presenting participants with a cue word (*dog*) and collecting the words that spontaneously came to their mind (*bark*, *pet*, ...) (Fig. 1, circles); (b) asked the same participants to *explain* their associations in a short sentence (Fig. 1, speech bubbles); and (c) labelled a subset of explanations with a relation type from a pre-defined set (Fig 1, arrow labels). The resulting data set is of high quality and diversity, albeit smaller in size tnan CNREL.

Using our data sets, we present a series of initial experiments. We test three models proposed in the recent unsupervised relation extraction (URE) literature. Results show the utility of our data sets, and that common sense relation extraction is a challenging task, constituting fruitful ground for future research on common sense knowledge induction. In sum, our contributions are

- The new task of common-sense relation extraction from natural language
- Two large-scale data sets, with different size and quality trade-offs, to train and evaluate common sense relation extraction models
- Experiments with three URE models adapted from the encyclopedic relation extraction domain, showing that broad-stroke common sense relations are learnt, and verifying the challenge of the task.

	Sentence [RELATION]
CNREL	a *bottle* is used to *hold* a liquid [USEDFOR]
	engine is part of *car* [PARTOF]
	you are likely to find *bread* in a *store* [ATLOCATION]
	bicycle racing is a sport [USEDFOR]
	army is used for *military* purposes [HASCONTEXT]
	wallet is about the same size as a *pocket* [LOCATION]
WAREL	*codes* are needed to *decipher* something. [FUNCTION]
	our *military* has a large *army* branch. [PARTOF]
	summer is always *hot*. [INHERENT-PROPERTY]
	the *leaves* started to fall in *autumn* [TIME]

Table 1: Example sentences encoding relation types, from CNREL (top) and WAREL (bottom). The concepts are highlighted in *blue*. The bottom three CNREL examples illustrate the noise in the data set.

2 Background

We describe the resources and paradigms underlying our own data sets, and previous work on URE.

2.1 `ConceptNet` and OMCS

The Open Mind Common Sense (OMCS)[3] (Singh et al., 2002) initiative was a decade-long effort to crowd-source natural sentences expressing common sense knowledge. A large portion consists of templated sentences, completed by crowd workers ('a *fork* is USED FOR ___'; see more examples in Table 1), later augmented with free-form crowd-sourced relation descriptions. `ConceptNet` (Speer et al., 2017) is one of the largest common sense KGs capturing general-domain knowledge, consisting of links between pairs of associated concept, labeled with one or more discrete relation types from an 'organically grown' relation ontology comprising 30 relation types (Liu and Singh, 2004).

`ConceptNet` was partially extracted from sentences in OMCS, leading to a natural alignment of concept pairs in `ConceptNet` with OMCS, and projection of relation labels to OMCS sentences.

2.2 Word Associations

Word associations (Deese, 1966; Kiss et al., 1973) are a prevalent paradigm in cognitive science to probe the human mental lexicon (Nelson et al.,

[3] `https://s3.amazonaws.com/conceptnet/downloads/2018/omcs-sentences-free.txt`

2004; Fitzpatrick, 2006). They reflect spontaneous human associations between concepts. In a typical study, a participant is presented with a cue word (*trombone*) and asked to spontaneously produce the words that come to mine in response (*music*, ...). Through large-scale crowd-sourcing studies covering over 12K cues and thousands of participants, a large word associations graph (`SWOW`; Deyne et al. (2019)) has been constructed, as a resource of human concept association strength. `SWOW` has recently been shown to be an effective knowledge resource for common sense reasoning models (Liu et al., 2021a). The *nature* of the underlying relations, however, is an open research problem.

2.3 Unsupervised Relation Extraction

Unsupervised relations extraction (URE) has been tackled predominantly in the context of factual relational knowledge about named entities. Typical models are presented with corpora of contexts mentioning pairs of entities and tasked with assigning inputs into clusters resembling the relations connecting concept pairs. Existing approaches can be grouped into generative and discriminative. Yao et al. (2011) extend the standard LDA model to URE by considering relations as topics and documents as co-occurred mentions along with the dependency features. In discriminative line, Marcheggiani and Titov (2016) propose to learn relation clusters using variational auto-encoder (VAE): the encoder is a relation classifier aiming to predict a relation for a given input, and the decoder reconstructs one entity given the predicted relation and the other entity. Follow-up work focused on stabilizing training (Simon et al., 2019), leveraged self-supervision via boostrapping (Hu et al., 2020), or developed better feature sets (Tran et al., 2020). The discriminative is advantageous as it allows to incorporate diverse relational representations, which is important in common sense domain. In this paper, we apply three recent URE models to common sense RE.

3 Common Sense Relation Extraction

3.1 Task Formulation

Our goal is to induce latent common sense relations between pairs of concepts from natural language text. As input, we assume a large corpus of sentences s which mention two concepts (c_1, c_2) of interest $D = \{(c_1, c_2, s)\}_1^N$ (see examples in speech bubbles in Fig 1 and Table 1). The task is to cluster these sentences into groups reflective of a ground-truth common sense relation (e.g., USED-FOR).

For unsupervised RE, we only require a large set of contexts, which are predictive of the relations of interest (rather than accidental co-mentions). For evaluation, we additionally require a smaller corpus, where sentences are labeled with the true relations. We present two such data sets below.

3.2 CNREL

We use distant supervision to derive a large-scale corpus of common sense relations holding between concept pairs from `ConceptNet` and OMCS. Specifically, following previous work on RE from Wikipedia (Lin and Pantel, 2001; Yao et al., 2011; Marcheggiani and Titov, 2016), we align a sentence s in OMCS with a relational triple (c_1, r, c_2) in `ConceptNet` (version 5.5;[4] Speer et al. (2017)) if both c_1 and c_2 are mentioned in s (exact string match based on the lemma); and label the sentence s with relation type r. Many aligned sentences will *not* be predictive of the relation (see Table 1). We enhance the quality of the data by filtering out triples using a list of criteria adapted prior work (Yao et al., 2012), with the intuition that in relation-relevant contexts, the two concepts should be mentioned close to one another and connected with semantically meaningful dependency path.[5]

Relation inventory The training set of CNREL covers all 30 `ConceptNet` relations, [6] (e.g, ISA, ATLOCATION, USEDFOR). For comparability with the WAREL data (§ 3.3), we include the 17 most common relations in the test and dev set.[7] We sampled up to 1K instances for each of the 17 most common relations, and split the resulting set into dev (20%) and test set (80%).

Summary Our final data set consists of 83K train, 3K dev and 11K test instances (details in Table 4

[4] `ConceptNet` and OMCS are open source, licensed under CC BY SA 4.0.

[5] We retain triples whose `ConceptNet` confidence score is > 1; filter out sentences of length > 30 words, sentences where the two concepts are < 10 words apart or the dependency path connecting the words is of length < 10. Finally, the dependency path (from `benepar` model in spaCy 3.0.6) must not contain the labels 'parataxis', 'pcomp', or 'punct'.

[6] For detailed definitions and examples see https://github.com/commonsense/conceptnet5/wiki/Relations

[7] The full set (and distribution) of 30 `ConceptNet` relations is in Fig. 6 (Appendix), and the 17 test relations and their distribution in Appendix Fig. 7.

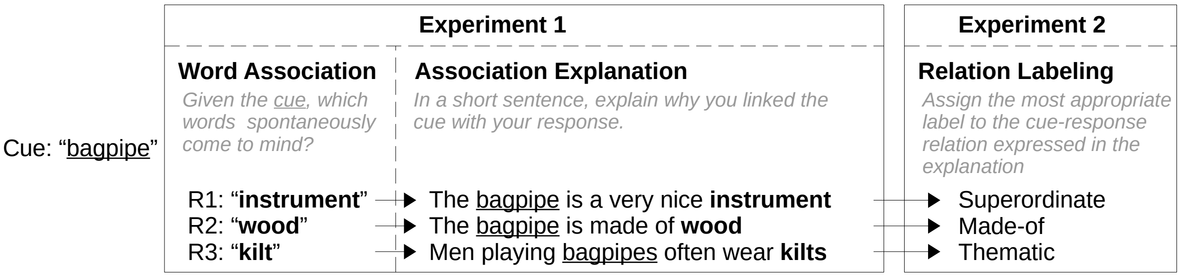

Figure 2: Overview over the data collection paradigm for WAREL.

in the Appendix). The heuristic alignment of CNREL allowed us to construct sizeable labelled dev and test sets. However, the relation labels remain noisy even after aggressive filtering. For instance, the example of *wallet* and *pocket* in Table 1 encodes the SIZE between the two concepts, instead of the intended LOCATION relation. Furthermore, sentences tend to be of a templated nature, calling into question the extensibility of models learnt on CNREL to other domains (e.g., corpora of news or web text). We address this question in our experiments (§ 5.4), and we propose a second data set which is of higher quality and diversity next.

3.3 WAREL

We propose a new framework to collect common sense relations between pairs of concepts (words) by crowd-sourcing explicit explanations of the relations. We adopt the word association paradigm (§ 2.2). Previous work (Liu et al., 2021a) has shown that large-scale word association network (WAN) contain common sense knowledge that can benefit common sense reasoning models for NLP. However, WANs typically provide responses associated with a cue word, while the underlying reasons or relations between cue-association pairs remain unknown. This lack of explainability limits its application to relation reasoning tasks. Our new data set can help to understand *why* humans make certain associations, and can serve as an explicit knowledge resource for reasoning models.

We collect the WAREL dataset by crowdsourcing via Amazon Mechanical Turk using a two-stage framework (Fig. 2). We first introduce our relation inventory, before describing the paradigm on a high level. Our study was approved by the university ethics board, and workers were paid above minimum wage. Detailed information is provided in Appendix A.

Relation Inventory The relation inventory underlying human word associations has been addressed on a theoretical or small-scale experimental level (Wu and Barsalou, 2009; McRae et al., 2012), and we construct a relation type inventory based on these works. We do not adopt `ConceptNet` relations, because (1) they resulted from the aggregation of several sources, baring a theoretical justification; (2) are dominated by overly broad types (HASCONTEXT); (3) contain several very similar types (CAUSES and HASSUBEVENT) that are hard to distinguish reliably in a crowd sourcing setup. Departing from the set of (Wu and Barsalou, 2009), we ran three pilot studies and converged on an inventory of 16 relations. The full set, including examples is presented in Fig. 8 and Table 7 in the Appendix.

Experiment 1 In the first experiment, we collect (a) word associations and (b) explanations from the same annotator, ensuring that the explanation indeed explains the intended, underlying association. Given a cue word, a worker first generates up to three spontaneous associations (Fig 2, left), and immediately after provides natural language explanations to describe why they linked the cue and each association (Fig 2, center). The resulting explanations will serve as our text corpus of sentences expressing relations between concept pairs.

The cue words in our experiment (N=1100) were sampled from a large-scale word association KG (`SWOW`; §2.2), ensuring a balanced distribution over the POS tags N, V, ADJ and ADV; as well as abstract vs concrete concepts. A single batch consisted of 5 randomly sampled cues, for which the worker provided associations and explanations. Each batch was labelled by 10 different workers.

Word associations and underlying reasoning are subjective, hence standard quality assessment via annotator agreement does not apply. Instead, we ensure high data quality by filtering responses wrt. a number of criteria including explanation length and diversity (cf., Appendix C.1 for details). We retained the annotations of 258 workers (out of 326). The final data set comprises 15K cue-association pairs along with 19K explanations.

Experiment 2 In a second experiment, we collected explicit relation labels for a subset of the annotations obtained in Experiment 1, as a development and test set for common sense relation extraction models (Fig 2, right). Given tuples of cue, association and explanation (c_1, c_2, s) a worker will choose the most appropriate relation type from the relation inventory explained above.

We sampled 757 instances from the data from Experiment 1 for labeling, excluding template-like explanations (e.g., "A is a B") to create a challenging test set and avoid the prevalence of template sentences characteristic of OMCS. The data includes cue POS-tags N, V and ADJ, as ADV associations proved challenging to annotate. We ensure high-quality labels through (a) detailed instructions; (b) a training phase; (c) careful selection of 45 reliable crowd workers who achieved accuracy > 0.5 in training; and (d) continuing feedback to annotators throughout annotation.

Each (c_1, c_2, s)-tuple was labeled by 5 workers. The ground truth was derived through majority voting, if the class was chosen by at least 3/5 workers. Otherwise, a label was chosen by one of the paper authors. We discard 53 instances for which none of the two workers agreed.[8] The final data set consists of 699 labeled instances, split into 50/50 test/dev.

Summary Our final dataset consists of 19K train, 350 dev, and 349 test instances. Unlike CNREL, this dataset conveys explicit relations between concepts , rather than accidental co-occurrences, and is of higher linguistic diversity. Furthermore, the WAREL dev and test set labels were manually verified by humans. Examples are provided in Table 1 (bottom). OMCS is the result of a decade-long collection effort, whereas WAREL was efficient to obtain via crowd-sourcing, and hence can be efficiently scaled up, or extended to other languages.

4 Relation Extraction Framework

In the remainder of the paper, we apply a series of recent models from the URE literature to the common sense domain, using our proposed data sets. We frame the task as open-domain relation discovery where no predefined relationships. Given a sentence s mentioning a pair of concepts c_1 and c_2, a URE model learns (1) to map the sentence to a latent relation representation ("encoder"); and (2) a relation classifier to assign the representation to a discrete relation cluster; (3) a "link predictor" which reconstructs the relational triple as an unsupervised training objective. We evaluate the extent to which induce clusters reflect the underlying classes in the data.

Encoder and Relation Classifier For a given triple (c_1, c_2, s), the relation classifier predicts the relational distribution of a relation latent representation encoded by an encoder:

$$\begin{aligned} z &= w^\top g_\theta(c_1, c_2, s) + b \\ p(r \mid z) &= \frac{\exp(z_r + b)}{\sum_{r'} \exp(z_{r'} + b)}, \end{aligned}$$

where g_θ is an encoder that maps (c_1, c_2, s) to a high-dimensional representation; and $w^\top \in \mathcal{R}^{d\times K}$ is the parameters of relation classifier, d denotes the dimension of the latent representation, K is the number of clusters (a pre-defined model parameter), and z_r the r^{th} element of z.

Link Predictor A good latent relation representation z should capture relevant contextual information and be capable of predicting missing context. Accordingly, the link predictor calculates the probability of predicting a missing concept given the predicted latent representation and one known concept (e.g., c_2):

$$p(c_1 \mid c_2, r) \propto \exp(\psi(c_1, r, c_2)) \qquad (1)$$

where ψ is an energy function. The model for $p(c_2 \mid c_1, r)$ is analogous. Following previous work (Marcheggiani and Titov, 2016; Simon et al., 2019), we use the combination of RESCAL and selectional preferences as the energy function:

$$\psi(c_1, r, c_2) = \underbrace{\mathbf{u}_{c_1}^T \mathcal{A}_r \mathbf{u}_{c_2}}_{\text{RESCAL}} + \underbrace{\mathbf{u}_{c_1}^T B_r + \mathbf{u}_{c_2}^T C_r}_{\text{Selectional Preferences}}$$

where $\mathbf{u}_{c_i}$ is the concept embedding of c_i learnt via the model, $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$ are model parameters, optimized to reconstruct the missing concept.

[8] See Table 5 in Appendix C.2 for examples with varying levels of annotator agreement.

4.1 Learning

The URE model jointly learns the relation classifier and link predictor by maximizing the joint probability of relation classifier and link predictor,

$$\sum_{r \in \mathcal{R}} p(r \mid x) \log p\left(c_1 \mid c_2, r\right) \log p\left(c_2 \mid c_1, r\right).$$

Unfortunately, $p(c_i \mid c_{-i}, r)$ in Eq (1) requires iterating over all potential concepts in the vocabulary, a very large set in the common sense domain. Instead of a multi-class (softmax) classifier, which would be infeasible, we train a binary (sigmoid) classifier to distinguish a positive triple (c_i, r, c_{-i}) from a set of sampled negative triples. Correspondingly, the link predictor can be approximated as follows:

$$\begin{aligned}\mathcal{L}_{\mathrm{LP}}=\underset{\substack{(c_1,c_2,s)\sim\chi \\ r\sim g_\theta(s)}}{\mathbb{E}}&\Big[-2\log\sigma\left(\psi\left(c_1,r,c_2\right)\right)\\&-\sum_{j=1}^{n}\underset{c'\sim\mathcal{E}}{\mathbb{E}}\left[\log\sigma\left(-\psi\left(c_1,r,c'\right)\right)\right]\\&-\sum_{j=1}^{n}\underset{c'\sim\mathcal{E}}{\mathbb{E}}\left[\log\sigma\left(-\psi\left(c',r,c_2\right)\right)\right]\Big]\end{aligned}$$

where σ is the sigmoid function, $c'{\sim}\mathcal{E}$ denotes sample negative concepts from the vocabulary and n is the number of negative samples. Following (Simon et al., 2019), we add two extra regularizers to stabilize model predictions by encouraging to predict a skewed relational distribution ($\mathcal{L}_{\mathrm{S}}$) per instance and uniform distribution over all instances per minibatch ($\mathcal{L}_{\mathrm{D}}$),

$$\begin{aligned}\mathcal{L}_{\mathrm{S}}&=-\mathbb{E}_{(c_1,c_2,s)\sim\chi}\,p(r|s)\log p(r|s)\\\mathcal{L}_{\mathrm{D}}&=\mathbb{E}_{r\sim g_\theta(s)}\left(q(r)\log q(r)\right),\end{aligned}$$

where $q(r)=\sum_{i=1}^{B}\frac{p(r|x_i)}{B}$ is the mean predicted relation within a minibatch of size B, leading to the final loss,

$$\mathcal{L}=\mathcal{L}_{\mathrm{LP}}+\alpha\mathcal{L}_{\mathrm{S}}+\beta\mathcal{L}_{\mathrm{D}}, \quad (2)$$

with α and β being hyper-parameters to control the strength of each regularizer.

Unsupervised Training In unsupervised training, the model is trained via Eqn (2). The labelled data is only used for model selection.

Supervised Training As our relation inventory is a set of closed relation types with limited numbers and is shared between dev and test, making it feasible to train a relation classifier using dev and compare the results with unsupervised training. We also include a supervised variant of the model, where we use a small amount of labelled data to train the relation classifier, and discard the link-predictor component. In this case, the loss is the cross-entropy between the gold and the predicted relation distribution:

$$\mathcal{L}_{\mathrm{CE}}=-\mathbb{E}_{(c_1,c_2,s)\sim\chi}\,y_r\log p(r|s),$$

where y_r is the true relation label.

5 Experiments

We instantiate the above framework with three encoders (explained below), and compare against a random baseline. We set $\alpha=0.01$, $\beta=0.02$ and $n=5$. For models trained and evaluated on in-domain data, we set the number of classes in the classifier same as the number of ground truth labels ($K=17$ for CNREL and $K=16$ for WAREL). For models evaluated on out-of-domain evaluation, we set the number of K as the combined of dev and test sets ($K=33$). All reported results are averages over three runs using different random seeds. Models are stable under runs, so we didn't report the variance.

5.1 Encoders

We conduct experiments with three types of encoders from the recent URE literature, which use different features.[9]

Feature (Marcheggiani and Titov, 2016) leverages 8 linguistic features to represent information covered in each input sentence and the entity pair, including the surface forms and POS tags of c_1 and c_2, and bag of words, POS sequence, and dependency path between c_1 and c_2, and the lemmas of trigger words from the dependency path. No parameters are learnt for the encoder function g, as all features are pre-defined.

EType+ (Tran et al., 2020) originally used entity type as information (Person, location, ...) as features, i.e., $g(c_1, c_2, s)=[\mathbf{c}_1^t, \mathbf{c}_2^t]$, where $\mathbf{c}_1^t$ and $\mathbf{c}_2^t$ indicate the entity type embeddings. In our experiment, we instead use the POS tag of entities, as

[9]We use the implementaions provided by Tran et al. (2020) https://github.com/ttthy/ure

c_1 and c_2 are not typically named entities in the common sense domain.

BERT embeds s using BERT (Devlin et al., 2019), and uses the concatenation of the final hidden layer of c_1 and c_2: $g(c_1, c_2, s) = [\mathbf{c}_1^b, \mathbf{c}_2^b]$. We use the BERT-base for all experiments, whose parameters are fixed during training.

5.2 Evaluations Metrics

We report results in terms of V-measure (Rosenberg and Hirschberg, 2007), an information theoretic measure of the extent to which clusters consist of instances from a single gold class (homogeneity), and to which all instance of a gold class are contained in a single cluster (homogeneity). V-measure is the harmonic mean of the two.

5.3 In-domain Results

Do recent models for encyclopedic relation extraction transfer to the common sense domain? We trained the models in § 4 separately on CNREL and WAREL and evaluated on the corresponding test sets. The left part of Table 2 presents the results. Note that the numbers are not comparable across CNREL and WAREL due to different evaluation sets and relation inventories. All models outperform the random baseline, and overall weak supervision (SRE) improved results (URE) even with a very small set of labels (N=350) for WAREL. BERT performs best in the unsupervised regime (URE), while Feature outperforms BERT under supervision. Supervision leads to larger improvements for CNREL than WAREL. This might be explained by the small WAREL development set.

5.4 Out-of-domain Results

An ideal common sense relation extraction model would be able to distill relations from *any* natural language resource. To this end, we apply models trained on WAREL to the “out of domain” CNREL data, and vice versa. Recall that the data sets differ both in style (CNREL being more templated) and relation inventory, constituting a challenging domain shift. Furthermore, we ask whether a model trained on a larger but noisier out-of-domain data (CNREL) has an advantage over a model trained on a smaller in-domain data set (WAREL). Models are trained and selected on the source domain and then tested on the target domain.

Results are shown in the right half of Table 2. Comparing with results in the left half, it can be seen that the transfer from CNREL to WAREL improved model performance across the board, while the transfer from WAREL to CNREL lead to performance degradation. This suggests CNREL has wider knowledge coverage, due to its larger scale. We further investigate the impact of training set size by training URE BERT on subsets of CNREL of varying size, and evaluating on WAREL. Fig. 3 shows that more data leads to higher performance, but also that URE BERT trained on an equivalent amount of in-domain WAREL data (a fifth of the size of CNREL) achieves higher performance (stars in Fig. 3). We conclude that high quality, in-domain data results in better performance when data scale is small, but this can be compensated with larger data scale.

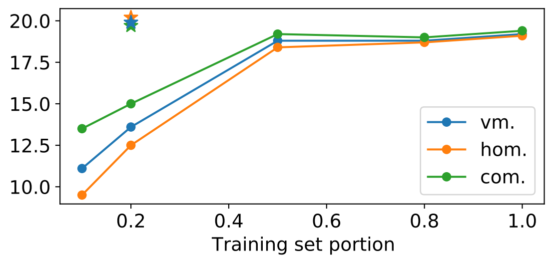

Figure 3: URE BERT trained on varying portions of CNREL train, and tested on WAREL. Stars show in-domain performance on the full WAREL (=0.2 × |CNREL |).

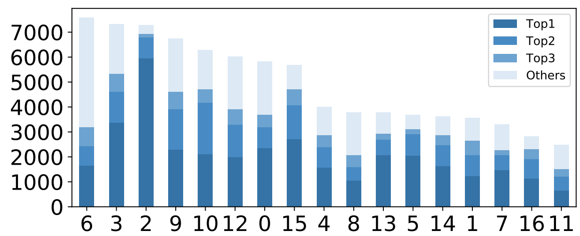

Figure 4: Relation clusters predicted by URE BERT on CNREL (in-domain). The x-axis is the cluster index. The y-axis is the number of instances per cluster. Top1–Top3 indicate the number of instances of the three most prevalent gold class labels.

5.5 Qualitative Results

We qualitatively inspect the clustering induced by the best-performing unsupervised model, namely in-domain BERT on CNREL. Following previous work (Yuan and Eldardiry, 2021), we measure the purity of each cluster by analysing its coverage of true relations. Ideally, each cluster would be dominated by a single (or few) gold class. Fig. 4 shows that most induced clusters are indeed dominated by

Test Set	Model	In-domain						Out-of-domain					
		URE			SRE			URE			SRE		
		vm	hom	com	vm	hom	com	vm	hom	com	vm	hom	com
CNREL	Random	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
	EType+	19.2	14.3	29.1	21.5	16.5	30.9	**19.9**	15.2	28.8	**14.8**	10.5	26.4
	Feature	20.4	19.5	21.5	**34.8**	33.8	35.9	12.8	10.1	20.2	8.5	6.4	12.8
	BERT	**23.4**	22.9	23.9	32.8	32.2	33.4	8.5	8	9	2	1.3	4.5
WAREL	Random	6.9	8.1	6.0	6.9	8.1	6.0	6.9	8.1	6.0	6.9	8.1	6.0
	EType+	13.2	9.6	21.6	11	7.2	25.2	16.8	13.4	22.9	20.7	17.7	24.9
	Feature	12.7	10	17.9	**26.8**	21.1	36.7	19	17.8	20.4	**21.4**	21	21.9
	BERT	**19.9**	20.2	19.7	18.5	14.1	27.2	**19.2**	19.1	19.4	20.3	20.2	20.4

Table 2: Common relation extraction results for models evaluated on CNREL (top) and WAREL (bottom). For In-domain results, models were trained on the training portion of the same data set. For, out-of-domain results models were trained on the respective other data set. We report homogeneity (hom), completeness (com) and V-Measure (vm), averaged over three runs.

C6	MANNEROF, CAUSES, ISA, HASSUBEVENT, HASPREREQUISITE
C2	ISA, ATLOCATION, HASA, PARTOF, HASPROPERTY
C4	DESIRES, NOTDESIRES, HASPROPERTY, ATLOCATION, CAPABLEOF
C12	USEDFOR, MANNEROF, ISA, CAPABLEOF, RECEIVESACTION

Table 3: Top five true relation labels in induced clusters 6, 2, 4, and 12 by BERT URE on CNREL.

the top three relation labels (but see e.g,. cluster 6 for an exception).

We print the top 5 dominating gold classes for selected clusters in Table 3. C6 covers action related relations, while C2 relates to the spatial and part-whole properties of objects. Desires/goals are captured in C4, while C12 covers 'utility' knowledge. Overall, we also observed that the most dominant relation in CNREL, ISA, penetrates most clusters. While overall, our results indicate that BERT learns broad-stroke common sense relations in an unsupervised manner, there is ample room for future work.

6 Discussion and Conclusion

We introduced the new task of common sense relation extraction from natural language corpora. We formalized the task as unsupervised clustering of sentences s which express a relation between two mentioned concepts c_1 and c_2, and contributed two data sets for model training and evaluation: The larger yet noisier CNREL, where sentences were heuristically aligned with concept/relation tuples and hence often do not reflect the underlying relation. WAREL is a crowd-sourced data set of word association explanations, ensuring that all sentences indeed express a relation between concepts. Initial experiments with existing relation extraction models under no or little supervision show that some meaningful relation clusters emerged, and that common sense RE is a challenging task, with ample scope for future work.

We adopted encoders from the encyclopedic domain, and one direction for future work would be the development of common-sense adapted sentence encoders, such as the pre-trained COMET model (Bosselut et al., 2019). Ample recent work has probed large pre-trained language models for common sense knowledge (Trinh and Le, 2018; Cui et al., 2021). This line of work can be extended to the more challenging common sense *relation* probing, using the high-quality WAREL data as a testbed. Finally, the WAREL sets could also be used to train and test models for common sense relation generation; and our resource of relational common sense knowledge can be incorporated into reasoning models for downstream tasks like question answering.

Our WAREL collection paradigm is efficient (it took <4 months compared to decades of effort for OMCS) and hence can be extended to other languages, communities and cultures. This provides the opportunity to collect diverse associations avoiding the pitfalls of a bias toward English-speaking cultures in NLP (Liu et al., 2021b).

References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Çelikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Leyang Cui, Sijie Cheng, Yu Wu, and Yue Zhang. 2021. On commonsense cues in BERT for solving commonsense tasks. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 683–693. Association for Computational Linguistics.

Ernest Davis and Gary F. Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58:92 – 103.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, Hong Kong, China. Association for Computational Linguistics.

James Deese. 1966. *The structure of associations in language and thought*.

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

S. Deyne, D. Navarro, Amy Perfors, M. Brysbaert, and G. Storms. 2019. The "small world of words" english word association norms for over 12,000 cue words. *Behavior Research Methods*, 51:987–1006.

T. Fitzpatrick. 2006. Habits and rabbits: word associations and the l2 lexicon. *Eurosla Yearbook*, 6:121–145.

Tianyu Gao, Xu Han, Keyue Qiu, Yuzhuo Bai, Zhiyu Xie, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2021. Manual evaluation matters: Reviewing test protocols of distantly supervised relation extraction. In *FINDINGS*.

Xuming Hu, Lijie Wen, Yusong Xu, Chenwei Zhang, and Philip Yu. 2020. SelfORE: Self-supervised relational feature learning for open relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3673–3682, Online. Association for Computational Linguistics.

G. R. Kiss, C. Armstrong, R. Milroy, and J. Piper. 1973. An associative thesaurus of english and its computer analysis. *The Computer and Literary Studies*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. Kagnet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of EMNLP-IJCNLP*.

Dekang Lin and P. Pantel. 2001. Dirt – discovery of inference rules from text.

Chunhua Liu, Trevor Cohn, and Lea Frermann. 2021a. Commonsense knowledge in word associations and ConceptNet. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 481–495, Online. Association for Computational Linguistics.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021b. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Liu and Push Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22:211–226.

Diego Marcheggiani and Ivan Titov. 2016. Discrete-state variational autoencoders for joint discovery and factorization of relations. *Transactions of the Association for Computational Linguistics*, 4:231–244.

Ken McRae, Saman Khalkhali, and Mary Hare. 2012. Semantic and associative relations in adolescents and young adults: Examining a tenuous dichotomy.

D. Nelson, C. McEvoy, and T. A. Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36:402–407.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP*.

Vered Shwartz, Peter West, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. Unsupervised commonsense question answering with self-talk. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4615–4629, Online. Association for Computational Linguistics.

Étienne Simon, Vincent Guigue, and Benjamin Piwowarski. 2019. Unsupervised information extraction: Regularizing discriminative approaches with relation distribution losses. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1378–1387, Florence, Italy. Association for Computational Linguistics.

Push Singh, Thomas Lin, Erik T. Mueller, Grace Lim, Travell Perkins, and Wan Li Zhu. 2002. Open mind common sense: Knowledge acquisition from the general public. In *OTM*.

R. Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

Thy Thy Tran, Phong Le, and Sophia Ananiadou. 2020. Revisiting unsupervised relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7498–7505, Online. Association for Computational Linguistics.

Trieu H. Trinh and Quoc V. Le. 2018. Do language models have common sense.

Ling Ling Wu and Lawrence W. Barsalou. 2009. Perceptual simulation in conceptual combination: evidence from property generation. *Acta psychologica*, 132 2:173–89.

Limin Yao, A. Haghighi, S. Riedel, and A. McCallum. 2011. Structured relation discovery using generative models. In *EMNLP*.

Limin Yao, Sebastian Riedel, and Andrew McCallum. 2012. Unsupervised relation discovery with sense disambiguation. In *ACL*.

Chenhan Yuan and Hoda Eldardiry. 2021. Unsupervised relation extraction: A variational autoencoder approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1929–1938, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Dataset Collection Details for WAREL

Our study received ethics approval (# 2021-22495-22206-5) from the university ethics review board.

Full Instructions We collect the WAREL dataset by crowdsourcing via Amazon Mechanical Turk. Figure 5 presents the annotation interface. The instruction page, includes (1) the Plain English Statement for this project, including what data will be collected, how the data will be processed and used (2) a consent form to inform workers the potential any risks so that workers can decide whether to work on this task. To avoid any potential confronting content, we removed profane words [10] before sampling cue seeds from SWOW for Experiment1.

The payment for both experiments is calculated based on the minimum wage salary in the country where the authors located in, which is much higher than the United States (the location of our annotators).

Task and Payment for Experiment 1 We take 5 words as a batch and assign it to 10 workers. Each worker first produces up to three responses for all five words, and then generates an explanation given each pair of associated words. Workers can skip cues (if their meaning is unknown) or provide fewer than three responses (if they cannot think of more). Each batch is paid with \$0.66 reward with extra bonus up to \$1, depending on the number of known cues, associations and explanations. This task takes approximately 5 minutes, as estimated by locally conducted pilot studies. Finally, we paid an average of \$1.48 per batch (estimated time =5 mins; hourly wage $\approx$ \$17.76) .

Task and Payment for Experiment 2 Each batch consists of 30 (c_1, c_2, s) triples. A worker will select the most appropriate relation label from a pre-specified list to each triple in the batch. This task takes approximately 15 mins to 30 mins, varying from different individuals. The amount of time is estimated by three pilot by the authors and volunteers who are college students from the university. Each batch is paid with \$1 reward with extra bonus up to \$8, depending on the annotation quality. We paid an average of 5.92 per batch (estimated time = 15 mins to 30 mins; hourly wage $\approx$ \$11.84 to \$23.68).

Data Privacy and Usage Our collected data does not include any personal information except the worker ID, which is a unique identifier for each AMT worker. To anonymize the data, we removed the worker ID in our published dataset. Our collected data will be publicized on for research purpose.

B Data Statistics for WAREL and CNREL

Table 4 presents the statistics of CNREL and WAREL. It can be seen that the two datasets share some similarities in terms of the number of relation

[10]https://www.cs.cmu.edu/ biglou/resources/bad-words.txt

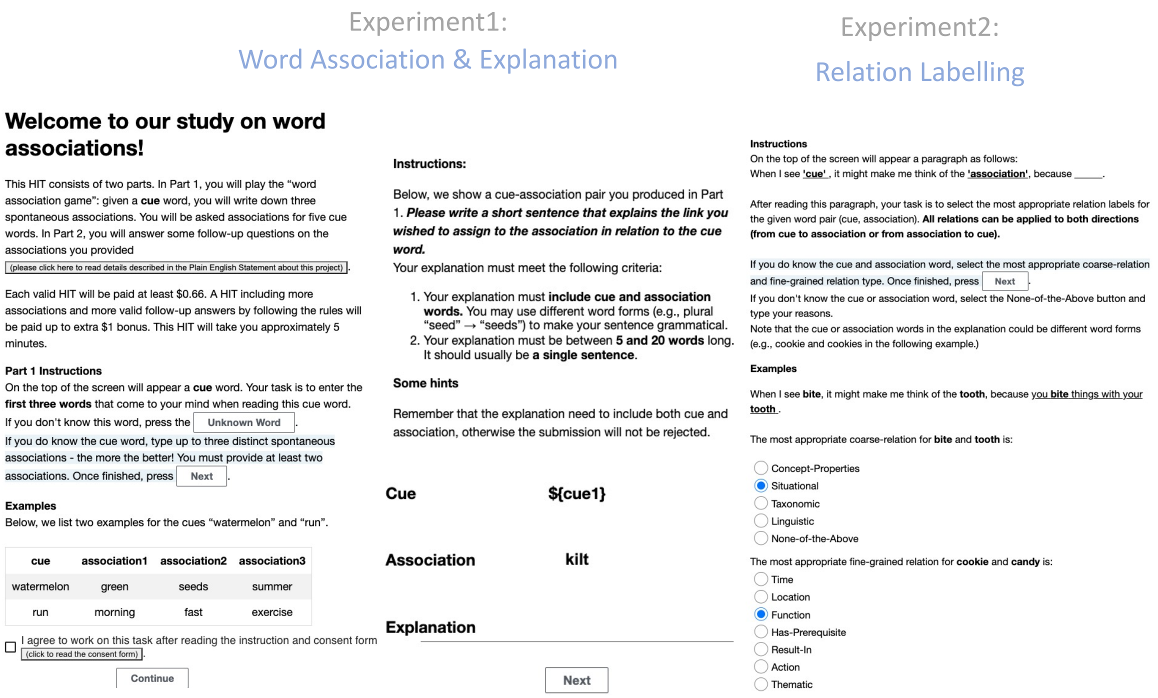

Figure 5: Annotation interface for collecting WAREL.

inventory, the average sentence length, the number of vocabularies. One key difference is the scale of sentences per (c_1, c_2) share. Each pair in CNREL are mentioned in multiple sentences (average is 6.7), but about only 1 sentence in WAREL.

B.1 CNREL

Figure 6 and Figure 7 present the train, dev and test relation distribution on CNREL. The dev and test set are both label balanced, but the distribution of the training set still have the long-tail problem.

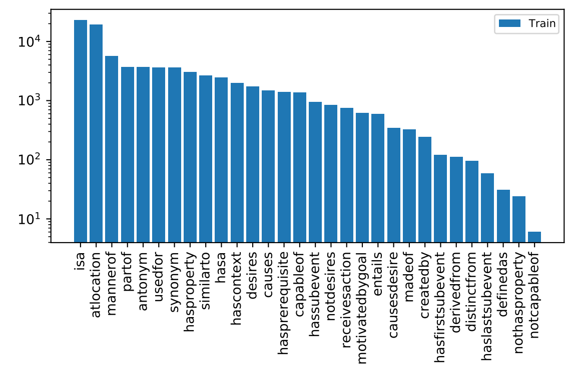

Figure 6: Relation distribution on CNREL training set.

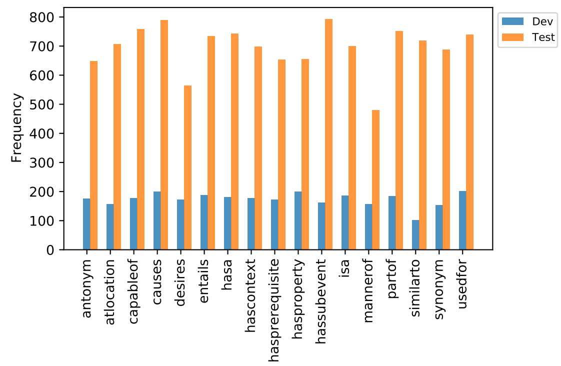

Figure 7: Relation distribution CNREL dev and test set.

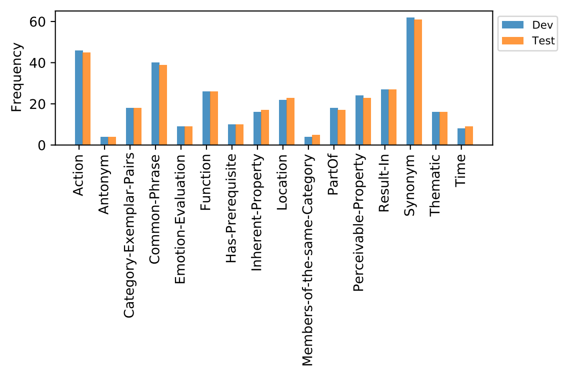

Figure 8: Relation distribution on WAREL dev and test set.

B.2 WAREL

Table 7 provides the definition for relation inventory we used for collecting WAREL. Figure 8

Dataset	split	#pairs	#sent	#cn_rel	#avg.sent_len	#vocab
CNREL	train	12342	83824	30	10.3	6470
	dev	1982	2956	17	10.5	1926
	test	4733	11826	17	10.4	3363
WAREL	train	15330	19002	-	9.7	6281
	dev	292	350	16	8.9	1225
	test	283	349	16	9.0	1220

Table 4: The statistics of CNREL and WAREL.

shows the relation distribution of WAREL dev and test set.

C Quality Control

C.1 Quality Control in Experiment 1

We list the detailed criteria to control the quality of the generated explanations in the Experiment 1. To collect high quality data, we introduce a number of strategies to control the quality, starting from the design of the guideline, to the selection of workers and the post-selection of explanations. In the guidelines, we set two criteria for the generated explanation: (1) the explanation must include the cue and association words. Different word forms (e.g., plural "seed" $\rightarrow$ "seeds") are allowed to ensure grammaticality; (2) the explanation should be a single sentence, and between 5 and 20 words long.

After obtaining the explanation, we filter out workers and explanations using the following criteria: a) workers who marked more than 3 of 5 cues as *unknown* b) workers whose explanations did not include the cue and association; or c) workers whose explanations rigidly follow a template (using manual inspection).

C.2 Labelled samples for Experiment 2

Table 5 presents some examples along their labels from five annotators. We discarded examples for which no two annotators agreed on a label (examples in the bottom part of Table 5).

D Training and Hyperparameters

All of our experiments are run on single GPU of NVIDIA V100 SXM2 (32G). As the parameters in encoder is small (or is fixed), the training time of each run is within an hour on both datasets.

Table 6 presents the hyperparameters we use in training three models. We manually tune the key hyper-parameter: learning rate on different sets using grid search from { 0.0001, 0.001, 0.005, 0.007} and the best one for each model is reported in Table 6.

Type	cue	association	explanation	{Annotation: count}
Retained	honey	sweet	honey is a very sweet substance.	{Perceivable-Property: 5}
	gypsy	europe	gypsies now mainly live in europe.	{Location: 4, Thematic: 1}
	baked	fried	baked and fried are two ways to prepare food.	{Members-of-the-same-Category: 3, Thematic: 1, PartOf: 1}
Discarded	buddy	together	buddies love to spend time together.	{Emotion-Evaluation: 1, Result-In: 1, Thematic: 1, Location: 1, Inherent-Property: 1}
	breath	oxygen	when you breath you inhale oxygen.	{Result-In: 1, Action: 1, PartOf: 1, Has-Prerequisite: 1, Function: 1}
	faithful	committed	being faithful in a relationship involves being committed to the other person.	{Members-of-the-same-Category: 1, Thematic: 1, PartOf: 1, Synonym: 1, Has-Prerequisite: 1}
	staff	employed	staff is the people employed by a particular organization.	{PartOf: 1, Thematic: 1, Has-Prerequisite: 1, Members-of-the-same-Category: 1, Function: 1}

Table 5: Samples of retained and discarded instances in WAREL Experiment 2. The Annotations column indicates the labels assigned to the instance together with assignment count out of 5 annotations.

Parameter	**Value**
Optimizer	AdaGrad
Number of epochs	10
Learning rate	0.007
Batch size	100
Feature dimension	10
Early stop patience	3
L_s coefficient	0.01
L_d coefficient	0.02

(a) Feature

Parameter	**Value**
Optimizer	Adam
Number of epochs	10
Learning rate	0.001
Batch size	100
Early stop patience	10
Entity type dimension	10
L_s coefficient	0.01
L_d coefficient	0.02

(b) EType+.

Parameter	**Value**
Optimizer	Adam
Number of epochs	5
Learning rate	0.001
Batch size	64
Early stop patience	3
L_s coefficient	0.01
L_d coefficient	0.02

(c) BERT

Table 6: Hyper-parameter values used in our experiments.

Coarse Relation	Fine-grained Relation	Definition
Concept-Properties	Perceivable-Property	A perceivable property, including shape, color, pattern, texture, size, touch, smell, and taste.
Concept-Properties	PartOf	A a part or component of an entity or event.
Concept-Properties	Inherent-Property	The inborn, native or instinctive properties, which cannot be directly perceived when encountering a concept, that requires some kind of inference from perceptual data.
Concept-Properties	Material-MadeOf	The material of something is made of.
Concept-Properties	Emotion-Evaluation	An affective/emotional state or evaluation toward the situation or one of its components.
Situational	Time	A time period associated with a situation or with one of its properties.
Situational	Location	A place where an entity can be found, or where people engage in an event or activity.
Situational	Function	The typical purpose, goal or role for which cue is used for association. Or the reverse way.
Situational	Has-Prerequisite	In order for the cue to happen, association needs to happen or exist; association is a dependency of cue. Or the reverse way.
Situational	Result-In	The cue causes or produces the association. Or the reverse way. A result (either cue or association) shoud be involved.
Situational	Action	An action that a participant (could be the cue, association or others) performs in a situation. Cue and association must be among the (participant, action, object).
Situational	Thematic	Cue and association participate in a common event or scenario. None of the other situational properties applies.
Taxonomic	Category-Exemplar-Pairs	The cue and association are on different levels in a taxonomy.
Taxonomic	Members-of-the-same-Category	The cue and assoiation are members of the same category.
Taxonomic	Synonym	The cue and associaiton are synonym.
Taxonomic	Antonym	The cue and association are antonym.
Linguistic	Lexical	Cue and association share the same base form.
Linguistic	Common-Phrase	The cue and association is a compound or multi-word expression or form a new concept with two words.
Linguistic	Sound-Similarity	The cue and association are similar in sound.
None-of-the-Above	None-of-the-Above	Use this label only if other labels can not be assigned to the instance or you don't understand the cue, association or explanation.

Table 7: The definition of associative relations used for labelling WAREL.