WAX: A New Dataset for Word Association eXplanations

Chunhua Liu[†] **Trevor Cohn**[†] **Simon De Deyne**[‡] **Lea Frermann** [†]
[†]School of Computing and Information Systems
[‡]Melbourne School of Psychological Sciences
The University of Melbourne
chunhua@student.unimelb.edu.au
{tcohn,simon.dedeyne,lfrermann}@unimelb.edu.au

Abstract

Word associations are among the most common paradigms for studying the human mental lexicon. While their structure and types of associations have been well studied, surprisingly little attention has been given to the question of *why* participants produce the observed associations. Answering this question would not only advance understanding of human cognition, but could also aid machines in learning and representing basic commonsense knowledge. This paper introduces a large, crowdsourced dataset of English word associations with explanations, labeled with high-level relation types. We present an analysis of the provided explanations, and design several tasks to probe to what extent current pre-trained language models capture the underlying relations. Our experiments show that models struggle to capture the diversity of human associations, suggesting WAX is a rich benchmark for commonsense modeling and generation.[1]

1 Introduction

Word associations (Deese, 1966; Kiss et al., 1973) are a prevalent paradigm in cognitive science for probing the human mental lexicon (Nelson et al., 2004; Fitzpatrick, 2006). They reflect spontaneous human associations between concepts. In a typical study, a participant is presented with a cue word (e.g., *bagpipe*) and asked to spontaneously produce the words that come to mind in response (*music*, ...). Through large-scale crowd-sourcing studies covering over 12K cues, 3M responses and thousands of participants, a large word association graph (SWOW; Deyne et al. (2019)) has been constructed, as a resource of basic human conceptual knowledge. This repository of shared associations can serve as a source of commonsense knowledge as shown recently by incorporating SWOW as knowledge resource into commonsense reasoning models (Liu et al., 2021).

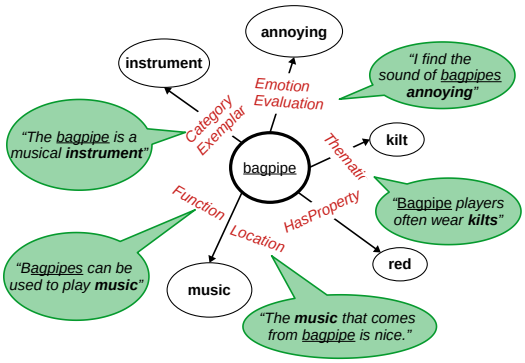

Figure 1: Excerpt of WAX, which consists of associations between cue words (bagpipe) and associations (kilt, red, ...) together with association explanations (speech bubbles) and discrete relation type labels (edge labels). Some associations are supported by distinct relation types and explanations (e.g., bagpipe→music).

However, existing word association data sets like SWOW only provide cue-association pairs, but do not further distinguish between different types of associations. To fill this gap, we constructed a novel data set to recover the underlying reasons by collecting associations together with free-text explanations from participants, and distill high-level relation types from them. Our data set can enhance our understanding of the *reasons* and *types* for conceptual associations in humans, and can serve as an explicit knowledge resource for reasoning models.

Our data set WAX (Word Association eXplanations) encodes English word associations with diverse explanations and high-level relation types and is illustrated in Figure 1. In a large crowd-sourcing study, we (a) collected human word associations by presenting participants with a cue word (*bagpipe*) and collecting the association words that spontaneously came to mind (*music*, *kilt*, ...) (Figure 1, circles); (b) asked the same participants to *explain* the link between the cue and

[1]Data and code are available at https://github.com/ChunhuaLiu596/WAX

Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 106–120
November 20–23, 2022.

their corresponding associations in a short sentence (Figure 1, speech bubbles); and (c) labeled explanations with a relation type adpated from a predefined set (McRae et al., 2012; Speer et al., 2017) (e.g., FUNCTION, edge labels in Figure 1). We ensure data quality through several layers of careful annotator training and data filtering.

Compared to existing work on categorizing word associations (Piermattéo et al., 2018; Fitzpatrick, 2006), WAX is larger in size, grounds associations in explanations, and will be released to the research community, supporting future research on understanding and modeling conceptual knowledge. WAX complements existing commonsense knowledge graphs, which either involved decades of manual work (ConceptNet; Speer et al. (2017)), rely on highly templated responses, limiting their ability to reflect the natural diversity in human associations (ATOMIC; Sap et al. (2019)); or only indirectly link concepts via a shared scene (CommonGen; Lin et al. (2020)). WAX results from a new, scalable method of collecting general commonsense knowledge, while maintaining both quality and diversity of associations and explanations, and can be cheaply extended to other languages.

We annotated a subset of WAX with high-level, discrete relation labels, enabling us to quantify the diversity of human mental relations, and to evaluate machine learning models in their ability to (a) distinguish different relations; and (b) generate plausible association explanations. Our experiments using pre-trained language models demonstrate the value of WAX as a rich and challenging data set for a variety of commonsense modeling and generation tasks. In sum, our main contributions are:

- A large data set of word associations with free-text explanations, providing the justification for the relation, and relation labels, which can support scalable studies of the human mental lexicon, and the development of models of relation extraction, commonsense knowledge and explanation generation.
- Extensive experiments demonstrating the utility of WAX for commonsense relation classification and explanation generation.
- Insights into the relative ease of predictability of different relation types, giving rise to future development of targeted models, as well as relation ontologies that are tailored to 'empirical' relations emerging from the data.

2 Background

Our work relates to several research lines, including word associations, commonsense knowledge graphs, and explainability.

Word Associations Word associations, as reflections of human mental lexica, have been studied extensively in psychology (Kiss et al., 1973). In early studies, word associations were predominantly collected on a small scale from homogeneous participants (Nelson et al., 2004; Kiss et al., 1973). Recently, crowd-sourcing has proved effective for collecting large-scale word association data sets in several languages, i.e., English (Kiss et al., 1973; Deyne et al., 2019), Dutch (Deyne and Storms, 2008) and Japanese (Joyce, 2005). Among them, SWOW (Deyne and Storms, 2008; Deyne et al., 2019) is the largest multi-lingual word association graph, covering 14 languages.[2] However, the graphs only include directed associations between words pairs, rendering the underlying reasons for association unknown.

Types of mental associations were previously studied in cognitive psychology (Read, 1993; Sinopalnikova, 2004; Fitzpatrick, 2006; Santos et al., 2011; Yokokawa et al., 2002). Previous work (Fitzpatrick, 2006; Piermattéo et al., 2018) showed that relations of word associations can be recovered by (1) asking subjects to *explain* (in words or in writing) the reasons for the produced association, then (2) inferring a relation based on the explanations. We follow the methodology from the above works both to recover the association reasons (see our method description in §3) and to label a subset of our word associations with relation types. In contrast with previous work, where collected data sets were small (e.g., 100 cues) and were not made available to the research community, we provide a large-scale data set by gathering explicit explanations and relation types, to encourage future work on automatic association inference and relation labeling.

Several relation type ontologies have been proposed (Cann et al., 2011; Estes et al., 2011; Fitzpatrick, 2006; Wu and Barsalou, 2009; Bolognesi et al., 2017), which typically distinguish four broad relation categories: taxonomic (*apple*, *pear*), situational (*airplane*, *travel*), properties (*sweater*, *comfortable*), and linguistic/form (*hobby*, *lobby*).

[2] https://smallworldofwords.org/en/project/home

McRae et al. (2012) build on the broad categories above, and refine them into a total of 28 subtypes, which we used as the basis for our own association labeling scheme (§3.2).

Commonsense Knowledge In word association graphs, cue words are typically surrounded by a rich set of associations (60 on average in SWOW) provided by multiple participants responding to the same cue. Naturally, those associations could be considered as shared, basic knowledge or a source of commonsense knowledge. Equipping machines with such resources has attracted substantial attention (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.

However, acquiring such commonsense knowledge is a challenge because it is vastly diverse and not often explicit in language, leading to data scarcity. Commonsense knowledge is typically collected either in free-text format (OMCS: Singh et al. (2002)) or structured databases (e.g., ConceptNet: Speer et al. (2017); ATOMIC: Sap et al. (2019)). Liu et al. (2021) showed that the associations in SWOW (i.e., without relation labels) bring comparable benefits as ConceptNet in commonsense question answering. Enhancing word associations with relations could increase its utility as a source of acquiring commonsense knowledge. Association explanations can also support research into *interpretable* commonsense reasoning.

Recently, pre-trained language models (PTLMs) were tested as commonsense repositories (Petroni et al., 2019; Shwartz and Choi, 2020; Bhargava and Ng, 2022) by probing the extent of commonsense knowledge encoded in PTLMs or using PTLMs to construct (or complete) commonsense knowledge graphs (Malaviya et al., 2020; Zhou et al., 2020). Integrating existing knowledge (free-text or structured) with PTLMs has been shown effective for improved machine reasoning (Wiegreffe et al., 2022; Moghimifar et al., 2021), and having machines explain why a certain association exists could bridge between structured and text representations. We explore association explanation in §5.

Explainable Commonsense Previous work used generated explanations to improve downstream task performance, e.g., on question answering (Shwartz and Choi, 2020) and natural language inference (Rajani et al., 2019). Less research has attempted to generate explanations to construct structured commonsense resources. Dognin et al. (2020) align ConceptNet with OMCS using heuristic rules and propose dual learning to transfer between a knowledge graph and free text. However, their language data is templated, and their dataset is not public. Other work has retrieved representative contexts from large corpora (Hendrickx et al., 2009), or used templates to construct sentences from triples (Petroni et al., 2019). In §5 we use WAX to generate explanations that reflect the naturalness and diversity of human explanations. Another related data set, CommonGen (Lin et al., 2020), consists of crowd-sourced, short sentences describing a scene that includes a given set of concepts (common objects and actions). CommonGen is designed to test machines' compositional ability, but relations between concepts are implicit in the description. Compared to their work, WAX is more explicit, eliciting concept associations from workers directly; more specific as each explanation focuses on a relation between an association pair; and more general (incl. adjectives, adverbs, and abstract concepts). WAX could hence be used to augment knowledge graphs like SWOW with relation labels, or free-text explanations.

3 The WAX Corpus

We present our two-stage framework for collecting word association relations between pairs of concepts (words) by crowd-sourcing explicit explanations of the relations (Figure 2). In Phase 1, we collect associations and free-text explanations to elicit the underlying reasoning. In Phase 2, we label a subset of (cue, association, explanation)-tuples (c, a, e)[3] with relation types r to characterize the inventory of common relation types. Appendix A contains details on annotator instructions and payment, as well as quality control.

3.1 Phase 1: Eliciting Explanations

In phase 1, we collect (a) word associations and (b) explanations from the same annotator, ensuring that the explanation indeed explains the true underlying association.[4] Following Deyne et al. (2019), given a cue word, a worker first generates up to

[3]Throughout the paper, we use c, a, e, r to denote cue, association, explanation and relation respectively.

[4]While we could have annotated existing word associations with explanations, this would require inference of another person's reasons for the association. To remove this confound we elicit associations and explanations from the same worker.

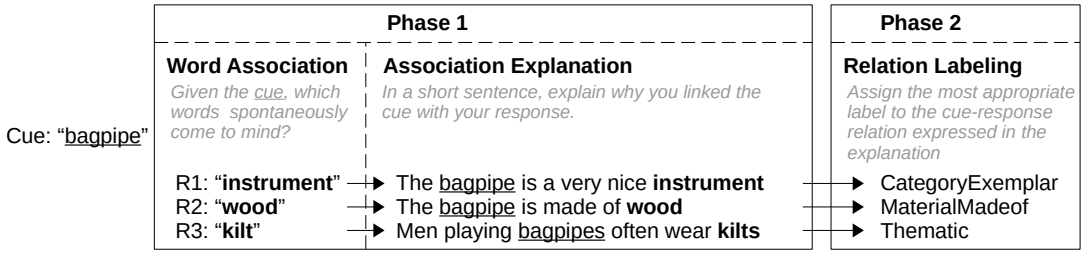

Figure 2: Overview over the data collection framework for WAX.

	Full WAX	Relation Labelled
# unique a	6,128	453
# unique (c, a)	15,337	520
# unique (c, a, e)	19,228	725
Vocab size	10,180	1,656
Avg len(e)	9.71	10.1

Table 1: The statistics of the full WAX, and its manually relation-labeled subset. Avg len(e) is the average explanation length (in words).

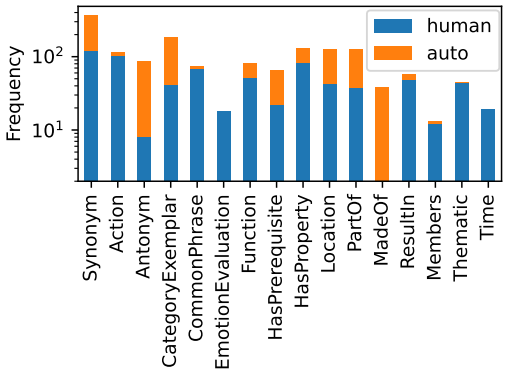

Figure 3: Relation distribution of WAX labeled data, including human labeled subset (bottom, blue), and auto-augmented subset (top, orange).

three spontaneous associations (Figure 2, left), and immediately after provides a one-sentence explanation of *why* they linked the cue and each association (Figure 2, center). The resulting explanations will serve as our text corpus of sentences expressing relations between concept pairs.

We used a set of 1,100 single-token cues, sampled from SWOW, ensuring a balanced distribution over the POS tags noun, verb, adjective and adverb; as well as abstract and concrete concepts. Each annotation batch consisted of 5 randomly sampled cues, each cue was labeled by 10 different workers on Amazon Mechanical Turk (MTurk). The final data set includes the annotations of 258 workers and comprises 15K unique cue-association pairs along with 19K explanations (Table 1, left).

3.2 Phase 2: Relation Labelling

Phase 2 augments the dataset above with explicit relation labels (Figure 2, right), as (a) a lens into the distribution of underlying association types; and (b) a testbed to examine machines' ability to extract or generate word association relations or explanations. Given a triple of cue, association and explanation (c, a, e), annotators choose the most appropriate relation type from a fixed relation inventory. We first introduce the relation inventory, before describing the process of relation labeling.

Relation Inventory We adapt an established semantic relatedness taxonomy of 28 relation types from cognitive studies of the human mental lexicon (Wu and Barsalou, 2009; McRae et al., 2012) and from ConceptNet (Speer et al., 2017). In multiple pilot annotations, we assessed the confusability and applicability of the relations to our association data. We conflated associations which were (i) similar (e.g., ACTION and BEHAVIOR), (ii) rare (e.g., ORIGIN), (iii) of opposite directionality (e.g., PARTOF and LARGERWHOLE), as this nuance was often not reflected in the explanations. The final label set consists of 16 relation types, which are listed in Figure 3 and, in more detail in Appendix A.1.

Relation labeling We sampled 757 instances from the data from Phase 1, excluding recurring template-like explanations (e.g., "A is a type of B") to create a challenging test set. We included cues with all POS from §3.1 except for adverbs.[5]

MTurk annotators were given the 16 relation types, their definitions, and examples. Each batch consisted of 30 (c, a, e) tuples, and a worker selected the most appropriate relation per tuple. Each batch was labeled by 5 workers and we derived

[5] Associations with adverbs have received little attention and are not well-covered by existing relation ontologies.

Criteria	WAX	Random
Q1: e valid explanation for (c, a)	0.98	-
Q2: r valid relation for (c, a)	0.79	0.26
Q3: r valid relation for (c, a, e)	0.76	0.20

Table 2: Manual validation accuracy for assessing explanations and their relation labels, as well as whether they are concordant with the cue and association pair. Also shown is the judged accuracy of instances with randomly corrupted relation labels.

gold labels for each (c, a, e) by majority vote.[6]

The final labeled data set consists of 725 (c, a, e)-tuples, covering 520 unique (c, a) pairs, labeled with one of 16 relations. The corresponding relation distribution is shown in Figure 3 (blue), showing that the relations are present in the data to varying degrees (e.g., the top 4 relations cover 52% of overall labeled data). Table 1 presents the full statistics of WAX. Examples are provided in Figure 1 and Tab 4. The collection of WAX was efficient (200 hours of crowd-sourcing), and hence can be scaled up, or extended to other languages.

3.3 Corpus Analysis

Quality In a final round of quality control, we examined the overall consistency of WAX. We designed three questions to manually examine its key elements: explanations, relations, and their alignment (Table 2). Q1 asks whether the generated explanation expressed a valid relation for the (c, a) pair. Q2 verifies the relation label quality by asking whether the given relation is valid for the (c, a) pair. Q3 looks into the alignment between explanations and relations by asking whether the explanation e indeed expresses the relation label r.[7]

We presented a random sample of 100 (c, a, e, r)-tuples from WAX to two qualified annotators[8] to answer the three questions. We additionally mixed in 50 (c, a, e) with a randomly assigned relation label r, as a reference point for random performance.[9] Table 2 shows the results. We can see that almost all explanations are a valid link between cue and association (Q1), demonstrating the validity of explanations from Phase 1. Close to 80% of relations are deemed valid for (c, a) (Q2) and (c, a, e) (Q3). To put this in perspective, the respective accuracy on the random sample were significantly lower. To the best of our knowledge, WAX is the first large-scale data set with explanations of conceptual associations.

Cluster	Representative TF/IDF 3-grams
LOCATION	'keep my in' 'my in my' 'put my in' 'on my face' 'many in my'
{SYNONYM, ANTONYM }	'the opposite of' 'opposite of is' 'is the opposite' 'is synonym for' 'another word for'
FUNCTION	'be used to' 'can be used' 'when you have' 'there is usually' 'in order to'
TIME	'am about something' 'if am about' 'if something will' 'something will happen'
ACTION	'in charge of' 'charge of the' 'was in charge' 'the helped the'
SIMILAR	'has similar meaning' 'similar meaning as' 'as has similar' 'meaning as has'
GENERIC1	'when you are' 'if you are' 'something you are' 'it when you'
GENERIC2	'referred to as' 'associated with being' 'think of as' 'in the past'
TOPICAL1	'in movie called' 'starred in movie' 'was in movie' 'books and movies'
TOPICAL2	'the game the' 'of the game' 'the ball in' 'to catch the' 'the game was' 'to win the'

Table 3: Representative sample of explanation clusters, as the top TF/IDF 3-grams. Clusters were labeled manually. Top: clusters aligning with predefined relations; center: topic-like clusters; bottom: generic clusters.

Clustering Explanations While classifying associative relations into a pre-defined ontology is an important task, both for comparability with prior cognitive work, and for model development and evaluation, it is informative to also group explanations in a purely data-driven way and compare the result against established relation inventories. To this end, we cluster all 19K WAX explanations using K-means in to 75 clusters.[10] In order to abstract away from signals specific to cue and association words, and focus on the general 'linking information', we masked cue and association tokens in the explanations and embedded the result with BERT-base (mean pooling over the final layer). We visualized each cluster by its top TFIDF trigrams.

Table 3 summarizes the clustering results. Some clusters capture relations in our ontology (LOCATION), although some relations are conflated

[6] Annotator agreement (pair-wise Cohen's κ) was $\kappa = 0.42$, indicating moderate agreement. 28 (c, a, e)-triples were removed, for which no majority could be reached.

[7] Table 8 (Appendix) shows examples for each question.

[8] One native speaker who was not involved in the project, and one of the authors.

[9] Note that the explanation for (c, a) was not randomized as this would have resulted in a trivial baseline.

[10] We experimented with smaller numbers of cluster but found that this number produced the most nuanced representations, and tried TFIDF instead of BERT embeddings which lead to highly skewed cluster memberships.

(*grater*, *cheese*) (1) "a grater is great to make shredded cheese."; (2) "he shredded the cheese with the grater"; (3) "i use a grater to grate cheese for my meal." (all FUNCTION)
(*flowing*, *water*) (1) "the water is flowing down the gutter."; (2) "water flows when you turn on the faucet."; (3) "water is often seen flowing through hills and valleys." (all ACTION)
(*reading*, *glasses*) (1) "he needs his reading glasses."; (2) "my father needs reading glasses."; (3) "the old man had to use reading glasses as it was difficult to see up close." (all COMMONPHRASE)
(*igloo*, *cold*) (1) "an igloo is very cold to the touch." (HASPROPERTY); (2) "the igloo is a cold place" (HASPROPERTY); (3) "when it's cold, you can build an igloo out of snow." (HASPREREQUISITE)
(*heaven*, *god*) (1) "heaven is where god lives." (LOCATION); (2) "heaven is the place where one can be with god." (LOCATION); (3) "it is said that heaven and hell were created by god." (ACTION)
(*goalie*, *save*) (1) "another job of the goalie is to save the shots on the goal" (FUNCTION); (2) "the goalie reached his glove out and made a big save" (ACTION)' (3) "the goalie had a great night, making a save on all but one of the shots he faced." (ACTION)

Table 4: Example WAX (c, a) pairs produced by >1 annotators, each with three explanations (1)–(3) and corresponding relation labels. The first three examples are *unambiguous* associations, where all explanations describe the same relation, while the last three are *ambiguous*, with explanations covering distinct relations.

(SYNONYM, ANTONYM). One general 'similarity'-focused cluster emerged, confirming previous findings on Enlgish native speakers' tendency to associate words based on general meaning similarity (Fitzpatrick, 2006). A second set of clusters captures 'generic associations' (GENERIC 1-2) such as 'If you are *c* then you *a*' or '*c* is associated with *a*'. The third (smallest) set is topical, with explanations referring to GAMES (sports) or ENTERTAINMENT (movies and music). Overall, we find that taxonomic and event-related (HASPREREQUISITE, RESULTIN) relations are well-captured, while property relations (MATERIALMADEOF, HASPROPERTY) are reflected to a lesser extent. This observation aligns with research showing that personal experiences (events and scenarios) inform word associations as well as conceptual representations more broadly (Barsalou, 1983).

Diversity Conceptual associations may result from factual knowledge, cultural or societal norms, or individual experiences. Here, we analyze the extent to which different annotators produced divergent associations and/or explanations (cf., the (*bagpipe* $\rightarrow$ *music*) association in Figure 1). The presented numbers are a lower bound on diversity, because WAX was collected from a small number of MTurk annotators, which were themselves not screened for diversity and are likely a homogeneous group of (western) English native speakers.[11]

15% (N=2358) of the (c, a) pairs in the full WAX[12] were produced by more than one annotator (3.5 times on average), raising the question whether a single typical relation or multiple distinct ones connect these concepts. We look into this by examining the labeled subset. For 59% (N=51) of these ambiguous (c, a) pairs, all annotators expressed the same underlying relation. Examples include (*grater*, *cheese*, FUNCTION), (*flowing*, *water*, ACTION) and (*reading*, *glasses*, COMMONPHRASE). For the remaining 41% (N=36) annotators expressed between 2 and 5 *different* relations. An example is the pair (*goalie*, *save*) produced by three annotators, with relations FUNCTION ($1\times$) and ACTION ($2\times$). Table 4 presents the above examples together with WAX explanations.

Analysis revealed that in cases where *different* relations emerged for the same (c, a) pair, these relations were predominantly event-related (HASPREREQUISITE, RESULTIN, ACTION, FUNCTION, CATEGORYEXEMPLAR). In §4 we explore the task of association relation classification, and evaluate our models on the challenging, ambiguous subsets described above to gauge the extent to which associative ambiguity is captured in different transformer-based classifiers.

4 Relation Classification

Automatic prediction of relation types or generation of explanations can support commonsense knowledge graph completion, enhance our understanding of such knowledge in pre-trained language models, or inform explainability research. In the following sections, we present a series of experiments to demonstrate how WAX can support progress towards some of these goals. This section addresses relation classification, before we study explanation generation in §5. We construct a relation classification task using our relation type ontology as ground truth, as a 16-way classification problem to predict a single relation type r

[11] We removed another layer of potential ambiguity in Phase 2, where we assigned a single label to each association by majority voting, even though some explanations may support several underlying relations.

[12] 16%(N=87) in the labeled proportion, accounting for 43% (N=312) of the labeled (c, a, e, r) tuples.

	Model	Overall (N=312)				Ambiguous relations (N=131)				Unambiguous relations (N=181)			
		P	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**	**Acc**
	Majority-Class	1.1	6.7	1.9	16.3	0.5	7.1	0.9	6.9	1.9	8.3	3.1	23.2
-EXP	LR	5.4	8.4	4.5	18.6	2.0	7.7	1.8	7.6	9.6	11.0	7.7	26.5
	BERT-base	24.8	26.8	20.7	32.8	23.9	23.2	18.8	26.2	22.6	25.1	21.0	37.6
	BART-large	34.5	48.0	35.9	47.8	30.9	35.5	29.4	38.2	37.4	42.8	37.3	54.7
+EXP	LR	29.9	17.7	16.0	22.1	23.1	14.5	10.9	16.0	32.3	16.5	16.1	26.5
	BERT-base	34.2	40.2	32.7	45.5	33.2	34.7	29.7	40.7	34.0	35.1	31.7	48.8
	BART-large	**49.6**	**57.7**	**48.1**	**56.2**	**41.9**	**47.2**	**37.7**	**48.9**	**47.2**	**50.5**	**46.1**	**61.5**

Table 5: Experimental results on relation classification, as macro precision, recall and F1, and accuracy for models with access to the full explanation (+EXP) or to cue and association only (-EXP). We report performance overall test instances (left), only relation-ambiguous (center), and only relation-unambiguous (right) instances.

from either only (c, a)-pairs (we call this model -EXP) or the full explanation e, which by construction includes c and a (+EXP).[13] We can thus test whether access to explanations, which lay out *why* two concepts are associated, improves relation prediction over and above the knowledge available to PTLMs via large-scale pre-training. For example, predicting a relation (e.g., FUNCTION) for the pair (*bagpipe*, *music*) is arguably simplified (or constrained) with access to an explicit explanation such as "*Bagpipes* are used to play *music*".

4.1 Dataset

As the labeled portion of WAX is both small in size and skewed in relation distribution (Figure 3), we augment its *training* portion with data from Wu and Barsalou (2009) and ConceptNet (Speer et al., 2017), which include concept pairs and their relation, but no explanations. To create labelled explanations, we find (c, a, r') edges in these external resources that are also in the unlabelled portion of WAX, (c, a, e), and then map the known relation label into our inventory, $r' \rightarrow r$, thus constructing full (c, a, e, r) tuples. In addition, we identified frequent patterns in the WAX explanations, and devised a small set of templates to extract the corresponding relations (e.g., 'a is part of c', indicates a PARTOF relation).[14] Those relations were verified independently by two authors of this paper, and we retained only instances where both agreed on their validity. We obtained 835 additional labeled explanations, as shown in Figure 3 (orange bars). The final data is split into 948, 300 and 312 (c, a, e, r)-tuples for train, dev and test sets, respectively.

[13] Another natural formulation is multi-class classification given as input a (c, a) pair with *all* produced explanations, which we leave for future work.

[14] All templates are shown in Appendix B.

4.2 Models

We experimented with discriminative and generative seq-to-seq methods for relation prediction. We fine-tuned BERT-base-cased (Devlin et al., 2019)[15] to embed the full explanation e (for explanation-aware models +EXP), or the simple template "c, [SEP], a" (for explanation-agnostic models -EXP); and use the hidden representation of the [CLS] token as input to a discriminative classification layer. In addition, we followed Huguet Cabot and Navigli (2021) and framed relation prediction as a sequence to sequence generation problem by generating (c, a, r) given (c, a, e) for +EXP, or given (c, a) for -EXP, using teacher forcing. While less direct, the approach is motivated by recent successes in formulating classical (structured) prediction problems as seq-to-seq (Bevilacqua et al., 2021; Nayak and Ng, 2020). Including c and a in the output lead to more focused r predictions, but also supports the prediction of entity-pair relations for explanations involving more than two entities. We fine-tuned BART-large (Lewis et al., 2020) as the generative model.[16] We compared against a logistic regression (LR) classifier with TF-IDF features, and a majority baseline. All models were trained on the training set, and hyper-parameters (Appendix C) were selected based on the dev set.

4.3 Results

Main results Table 5 (left block) presents the results. The fine-tuned LMs outperform the baseline models by a large margin, and BART per-

[15] It outperformed other BERT versions, incl. BERT-large.

[16] We represent the encoder input as "e <subj> c POS_c <obj> a POS_a", and the decoder input (with teacher forcing at training time) as "<triplet> c <subj> a <obj> r". <...> are sentinel token, and POS_x the POS tag of argument x. We use the code base from https://github.com/Babelscape/rebel.

(c, a)	Synonym	Prerequisite	Antonym	MadeOf	Location	PartOf	Function	ResultIn	Property	Emo-Eval	Time	Phrase	Action	Thematic	CategoryEx	SameMemb
darkness-light			⊗													
pocket-wallet					⊗											
skunk-smell									⊗	×						
printer-ink		×				○	×									
casino-money					⊗		⊗									
contact-phone	×				○		⊗						×			
lesson-learn		⊗			○			×					⊗			
discuss-talk	⊗	○					×						×		○	

Table 6: Selected relation classification results on unambiguous (top) and ambiguous WAX test instances, where each row shows the types of true (○) and predicted (×) relations when applied to the explanations for a cue-association pair.

forms better than BERT, suggesting the promising direction of modeling word association relations with seq-to-seq frameworks. We further explore this direction in §5. +EXP models (fine-tuned on full explanations) performed substantially better than -EXP models (fine-tuned on (c, a) pairs with no context), suggesting that explanations provide signal over and above the knowledge already encoded in PTLMs. This is confirmed by comparing against a BERT zero-shot model, which consistently performed worse than the majority class baseline (Overall accuracy of 5.6%). A class-wise performance analysis of the best model BART revealed that it was accurate for taxonomic relations and well-defined attributes (e.g., {SYNONYM, ANTONYM, PARTOF, LOCATION }), which are well-established in the literature, while situational associations (e.g., RESULTIN, HASPREREQUISITE) are not captured by the -EXP model, but are predicted at much higher quality by +EXP. Full details are in Appendix D. This concurs with the open challenge of event representations in NLP (Sap et al., 2019) and points to future work on tailoring models and relation sets. We estimate human accuracy at 76-79% (Table 2), leaving a substantial gap between model and human performance to be addressed in future work.

Relation diversity We evaluated our models separately on two challenging data subsets to investigate whether models capture the relation diversity discussed in §3.3: (1) (c, a) pairs with *multiple* explanations that all refer to the *same* relation type (Table 5, right block); and (2) (c, a) pairs with *multiple* relations that refer to *different* relation types (Table 5, center block). Transformer-based models outperform LR, with BART performing best. The difference between BART +EXP vs BART -EXP increases compared to Overall for both F1 and Acc, confirming the value of explicit explanations for these challenging subsets. Unsurprisingly, the ambiguous relation scenario is the most challenging.

We further analyze how model predictions differ from human labels on both relation-ambiguous and unambiguous (c, a) pairs. We inspect predicted labels from the best-performing model BART. Table 6 shows representative examples comparing human and model-predicted relations for unambiguous instances (one true relation, top) and ambiguous ones (multiple true relations, bottom). Although predictions diverge from gold labels, especially for the challenging ambiguous subset, the model labels are often reasonable. Consider (*discuss*, *talk*) with the explanation "to *discuss* something you must *talk* about it" and gold label CATEGORYEXEMPLAR, was predicted by the model as HASPREREQUISITE. It is not uncommon that taxonomic (CATEGORYEXEMPLAR) and associative or situational associations (HASPREREQUISITE, ACTION) relations are both valid for an explanation (Santos et al., 2011), leading to confusions by both our human annotators and model predictions. Our raw relation annotations include at least 5 annotations per (c, a, e) tuple, and hence capture this ambiguity which can be leveraged for model development and evaluation in future work. [17]

5 Generating Relation Explanations

Natural language inference or commonsense reasoning is often framed as mapping a free text input (e.g., a paragraph) to a structured output (e.g., a relation, (c, a, r) triple, or a multiple-choice answer). The underlying reasoning steps typically remain obscure. Constructing intuitive and faithful explanations for model predictions is an active research area of increasing impact. Mapping structured representations to natural language explanations is one approach, which has been limited by a lack of suitable training data sets. WAX is a parallel data set of structured relational data, aligned with diverse, human-generated free text explanations. Here, we show that it can support models to generate explanations which capture the diversity of human

[17] Our analysis also raises the question of how well the predefined relation ontology captures the relations encoded in the explanations. We clustered the explanations and observed it broadly aligns with our our relation ontology. See more details in Appendix 3.3.

	Prompt Relation	Generated Explanation
(a)	PARTOF	cowgirls wear boots as part of their outfits
	EMOTION	if someone is weeping they are likely sad.
	PROPERTY	lights are very bright when they are on
	PROPERTY	a shark has teeth that are very sharp
	RESULTIN	a hurricane can cause a major disaster in a community.
	ACTION	when trying to decipher a code, you need to first decipher the words
(b)	FUNCTION	i use money to make a payment for my car
	HASPREREQ	when you want to make a payment, you need to make money
	FUNCTION	casinos make money by selling alcoholic drinks
	LOCATION	casinos have lots of money
(c)	HASPREREQ	money is needed to operate a casino.
	RESULTIN	angles can bend in a triangle.
	FUNCTION	angles can be used to make triangle
	MADEOF	i prefer my light that is made of very bright glass
	LOCATION	water is flowing in a stream
(d)	TIME	water is a river that is flowing
	CATEXEMP	baked goods are a type of baked goods.
	EMOTION	i like to clench my fist when i am angry

Table 7: Illustrative examples of BART generated explanations in response to relation prompts of the form "c and a have a r relation." For each example, r is shown on the left and c and a are underlined in the generated explanation. Outputs are grouped to illustrate: (a) general quality, (b) diversity in generation for same (c, a) with ambiguous relations, and (c,d) unseen relation types with (c) plausible versus (d) nonsensical outputs.

reasoning. We fine-tune a generative PTLM to generate e given (c, a, r), noting that other tasks definitions are conceivable, including jointly generating structured predictions and explanations, e.g., predict (r, e) from (c, a).

5.1 Prompting Relation Explanations

Most relatedly, BART has been used to generate relational triples from sentences (Huguet Cabot and Navigli, 2021). Here, we investigate the more challenging, reverse, direction: generate a free-text explanation from a given (c, a, r)-triple encoded into the sentence prompt "c and a have a r relation". The output is a short sentence supporting the prompt. For example, the input "*bucket* and *wash* have a *function* relation", could elicit the output "I use a bucket to wash my car".

Setup Similar to §4.1, we augment the labeled training portion of WAX to increase its size and balance: for each (c, a, e, r) instance in the training data, we mask either c or a in the explanation and fill the blank with the top 10 candidates generated by BERT-large.[18] We down-sampled generated instances of overrepresented relation types, resulting in a balanced dataset of 12K (c, a, e, r) tuples, which are used to fine-tune BART. The original validation data is used for model selection. Table 11 (Appendix) lists the key hyper-parameters.

We explored the model explanations under four conditions: (a) prompting with human created (c, a, r)-triples from WAX (*dog, bark,* ACTION); (b) a version of (a) focused on ambiguous (c, a)-pairs, e.g., (*dog, guard,* ACTION) and (*dog, guard,* FUNCTION); (c) prompted as in (a) but with a relation *unseen* in WAX. These triples are often nonsensical (*dog, bark,* SYNONYM).

Results Qualitative results in Table 7 show that (a) explanations are overall relevant, factual and of high quality; (b) using nucleus sampling (Holtzman et al., 2020), we can generate different meaningful explanations for the same prompt; (c) the high quality extends to inputs that were not seen in WAX; and (d) for nonsensical triples, the model can still link the concepts with the given relation (2 and 3 in (d)) possibly leading to tautological outputs; or ignoring of the relation (1 in (d)). Our analyses suggest that WAX can be used for fine-tuning and probing commonsense knowledge in PTLMs, support future research into explanation generation, or bridging structured and free-text commonsense representations. We leave development of a quantitative benchmark to future work.

6 Conclusion

Word associations have been used as a lens into human conceptual representations for a long time, however, the *types* and *reasons* of these associations have not been studied at scale. We presented WAX, a large data set of word associations with explanations and relation labels. WAX is both an opportunity better understand the human mental lexicon, and a repository of relational commonsense knowledge both structured as (c, a, r) tuples, and free-text through the associated explanations. We demonstrated the utility of WAX for supervised relation classification and explanation generation; and presented a detailed data set analysis including association diversity and data-driven relation types. In future work, we plan to use WAX in tasks such as automatically labelling edges in commonsense knowledge graphs, commonsense question answering, and natural language inference.

[18] We inspected a sample of 80 prompts for validity.

Ethical considerations

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

Limitations We acknowledge that our dataset is collected from a limited number of English native speakers, and it can serve as an initial work to understand the underlying associative reasons *within this group*. Caution should be exercised when drawing general conclusions about human conceptual knowledge, and an important direction for future work is an extension to other languages. Reasons for associations are likely more diverse than reflected in our data set.

Data Privacy and Usage Our collected data does not include any personal information except the worker ID, which we redact from the data set. Our collected data will be made public for research purposes.

Acknowledgments

This work was supported in part by China Scholarship Council (CSC). We thank the reviewers for their valuable comments. This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200.

References

Lawrence W Barsalou. 1983. Ad hoc categories. *Memory & cognition*, 11(3):211–227.

Michele Bevilacqua, Rexhina Blloshmi, and Roberto Navigli. 2021. One spring to rule them both: Symmetric amr semantic parsing and generation without a complex pipeline. volume 35, pages 12564–12573.

Prajjwal Bhargava and Vincent Ng. 2022. Commonsense knowledge reasoning and generation with pre-trained language models: A survey. *CoRR*, abs/2201.12438.

Marianna Bolognesi, Roosmaryn Pilgram, and R Van den Heerik. 2017. Reliability in content analysis: The case of semantic feature norms classification. *Behavior Research Methods*, 49:1984–2001.

David R Cann, Ken McRae, and Albert N. Katz. 2011. False recall in the deese–roediger–mcdermott paradigm: The roles of gist and associative strength. *Quarterly Journal of Experimental Psychology*, 64:1515 – 1542.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

James Deese. 1966. *The structure of associations in language and thought. Baltimore: The Johns Hopkins Press*.

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. Association for Computational Linguistics.

Simon De Deyne, Daniel J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "small world of words" english word association norms for over 12,000 cue words. *Behavior Research Methods*, 51:987–1006.

Simon De Deyne and Gert Storms. 2008. Word associations: Norms for 1,424 dutch words in a continuous task. *Behavior Research Methods*, 40:198–205.

Pierre Dognin, Igor Melnyk, Inkit Padhi, Cicero Nogueira dos Santos, and Payel Das. 2020. DualTKB: A Dual Learning Bridge between Text and Knowledge Base. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8605–8616, Online. Association for Computational Linguistics.

Zachary Estes, Sabrina Golonka, and Lara L Jones. 2011. Thematic thinking: The apprehension and consequences of thematic relations. In *Psychology of learning and motivation*, volume 54, pages 249–294. Elsevier.

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

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 94–99, Boulder, Colorado. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language

generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Terry Joyce. 2005. Constructing a large-scale database of japanese word associations. *Glottometrics*, 10:82–99.

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*, pages 153–165.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Chunhua Liu, Trevor Cohn, and Lea Frermann. 2021. 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.

Hugo Liu and Push Singh. 2004. Conceptnet — a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Chaitanya Malaviya, Chandra Bhagavatula, Antoine Bosselut, and Yejin Choi. 2020. Commonsense knowledge base completion with structural and semantic context. volume 34, pages 2925–2933.

Ken McRae, Saman Khalkhali, and Mary Hare. 2012. Semantic and associative relations in adolescents and young adults: Examining a tenuous dichotomy. pages 39–66. American Psychological Association.

Farhad Moghimifar, Lizhen Qu, Terry Yue Zhuo, Gholamreza Haffari, and Mahsa Baktashmotlagh. 2021. Neural-symbolic commonsense reasoner with relation predictors. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 797–802, Online. Association for Computational Linguistics.

Tapas Nayak and Hwee Tou Ng. 2020. Effective modeling of encoder-decoder architecture for joint entity and relation extraction. volume 34, pages 8528–8535.

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.

Anthony Piermattéo, Jean-Louis Tavani, and Grégory Lo Monaco. 2018. Improving the study of social representations through word associations: Validation of semantic contextualization. *Field Methods*, 30(4):329–344.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! leveraging language models for commonsense reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

John Read. 1993. The development of a new measure of l2 vocabulary knowledge. *Language Testing*, 10:355 – 371.

Ava Santos, Sergio E. Chaigneau, W. Kyle Simmons, and Lawrence W. Barsalou. 2011. Property generation reflects word association and situated simulation. *Language and Cognition*, 3(1):83–119.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:3027–3035.

Vered Shwartz and Yejin Choi. 2020. Do neural language models overcome reporting bias? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6863–6870, Barcelona, Spain (Online). International Committee on 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 *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1223–1237, Berlin, Heidelberg. Springer Berlin Heidelberg.

Anna Sinopalnikova. 2004. Word association thesaurus as a resource for building wordnet. In *Proceedings of the Second International WordNet Conference, GWC 2004*, pages 199–205, Brno, Czech Republic. Masaryk University, Brno.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. volume 31.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2022. Reframing human-ai collaboration for generating free-text explanations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online and Seattle, USA. Association for Computational Linguistics.

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

Hirokazu Yokokawa, Satoshi Yabuuchi, Shuhei Kadota, Yoshiko Nakanishi, and Tadashi Noro. 2002. Lexical networks in l2 mental lexicon: Evidence from a word-association task for japanese efl learners. *Language education & technology*, 39:21–39.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. volume 34, pages 9733–9740.

A Dataset Collection Details for WAX

Our study received ethics approval with the application reference number of 2021-22495-22206-5 from the The University of Melbourne ethics review board.

We collect the WAX dataset by crowdsourcing via Amazon Mechanical Turk. Participants were informed what data will be collected, how the data will be processed and used, and asked for their explicit consent. To avoid potential confronting content, we removed profane words[19] before sampling cue seeds in Phase 1 (§3.1). The payment for both experiments was calculated based on the minimum wage in the authors' home country, which is higher than that of our workers.

Phase 1 collects word associations and corresponding explanations. Next we describe the collection details.

HIT and Payment Each batch (of 5 cue words) is assigned to 10 workers. Each worker (1) produces up to three associated words for each cue, and (2) writes an explanation for each association. 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 in a pilot study. We paid an average of $1.48 per batch, resulting in an hourly wage of $17.76 (all amounts in US dollars).

Quality Control Word associations and underlying reasoning are subjective, hence standard quality assessment via annotator agreement does not apply. Instead, we introduced a number of strategies to control quality: clear guidelines,[20] careful selection of workers, and filtering of explanations. A valid explanation must (1) include the cue and association words, or a morphological variant (e.g., plural); (2) be a single sentence of 5 to 20 words. We removed explanations which did not meet the criteria above or follow trivial templates, and batches where more than 3 of the 5 cues were marked *unknown*.

Phase 2 labels explanations with relations. Next we describe the HIT design and quality control.

HIT and Payment Each batch of 30 (c, a, e) triples is assigned to five workers. For each triple, workers select the most appropriate relation label from a given list (see Table 9 for list of labels and definitions provided to the workers). This task takes approximately 22 minutes, based on a pilot study. Each batch is paid at a minimum $1 with extra bonus up to $8, depending on the annotation quality. We paid an average of $5.92 per batch, resulting in an hourly wage of $17.36.

Quality Control We ensure high quality through (a) detailed instructions; (b) a training phase; (c) selection of 10 reliable crowd workers who achieved accuracy > 0.5 in training; (d) continuing feedback to annotators throughout annotation; (e) collecting labels from five workers for each (c, a, e). If a label has 3 or more votes it is selected; otherwise the instance is labeled by two experts (authors of the paper), and the voting test is re-applied.[21] We obtained an annotator agreement (pair-wise Cohen's κ) of $\kappa = 0.42$, indicating moderate agreement.

Final quality check Table 8 illustrates the questions used in our final WAX quality check, as described in Section 3.3 in the main paper.

Questions and Examples
Q1: Does the explanation express a valid reason for associating (c, a)? Example: raspberries can be made into jam.
Q2: Does the relation label express a valid relation for (c, a)? Example: (nature, beautiful, hasproperty)
Q3: Does the relation label express the relation for (c, a) that is described in the explanation? Example: (space, stars, partof, space has a lot of stars in it.)

Table 8: Examples of dataset quality check.

A.1 Relation inventory

Table 9 displays the relation ontology used in phase 2 of data collection, including a definition of each relation as presented to the crowd workers.

B Relation Templates

Table 10 lists trigger words and phrases used to automatically map recurring, templated WAX explanations to relations.

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

[20] The full guidelines will be released as part of the dataset.

[21] After this, 32 instances are still not assigned a label with three votes, and are discarded.

Broad Category	Relation	Definition
Concept-Properties	HASPROPERTY	Cue has association as a property; or the reverse. Possible properties include shape, color, pattern, texture, size, touch, smell, and taste; or inborn, native or instinctive properties.
	PARTOF	A part or component of an entity or event.
	MATERIALMADEOF	The material of something is made of.
	EMOTIONEVALUATION	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.
	LOCATION	A place where an entity can be found, or where people engage in an event or activity.
	FUNCTION	The typical purpose, goal or role for which cue is used for association. Or the reverse way.
	HASPREREQUISITE	In order for the cue to happen, association needs to happen or exist; association is a dependency of cue. Or the reverse way.
	RESULTIN	The cue causes or produces the association. Or the reverse way. A result (either cue or association) shoud be involved.
	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).
	THEMATIC	Cue and association participate in a common event or scenario. None of the other situational properties applies.
Taxonomic	CATEGORYEXEMPLAR	The cue and association are on different levels in a taxonomy.
	SAMECATEGORY	The cue and association are members of the same category.
	SYNONYM	The cue and association are synonyms.
	ANTONYM	The cue and association are antonyms.
Linguistic	COMMONPHRASE	The cue and association is a compound or multi-word expression or form a new concept with two words.
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 9: The definition of associative relations used for labelling WAX.

Relation	Trigger phrase
ANTONYM	opposite
PARTOF	part of
FUNCTION	used
CATEGORYEXEMPLAR	type of, form of
HASPREREQUISITE	require, need to
MATERIALMADEOF	make of/by/with
LOCATION	grow on, grown in, live in, on the, find
SYNONYM	similar, synonym, another word, define

Table 10: Templates used to automatically label explanations. Trigger word is the text between cue and association in the explanation.

	Classification		**Generation**
	BERT	**BART**	**BART**
Optimizer	AdamW_hf	AdamW	AdamW
Max Steps	500	1000	2000
Learning Rate	5E-05	2E-05	2E-05
Batch Size	8	8	4

Table 11: Experimental hyper-parameters.

C Hyperparameters

Table 11 lists the core hyperparameters used in the relation classification and generation experiments.

D BART class-wise relation prediction performance

Table 12 shows the class-wise relation classification performance of BART when fine-tuned on minimal templates (-EXP) and on full explanations (+EXP). The final column indicates whether access to explanations improved performance.

	Relation	BART -EXP			BART +EXP			
		P	R	F1	P	R	F1	Δ F1
(a)	SYNONYM	100.0	83.3	90.9	77.1	72.6	74.8	↓
	ANTONYM	100.0	100.0	100.0	75.0	100.0	85.7	↓
	ACTION	84.6	61.1	71.0	85.7	55.6	67.4	↓
	PARTOF	55.0	100.0	71.0	100.0	33.3	50.0	↓
	EMOTIONEVALUATION	50.0	100.0	66.7	42.9	60.0	50.0	↓
(b)	LOCATION	76.9	71.4	74.1	69.7	85.2	76.7	↑
	TIME	27.3	100.0	42.9	33.3	100.0	50.0	↑
	FUNCTION	23.5	26.7	25.0	63.6	48.3	54.9	↑
	HASPROPERTY	70.0	38.9	50.0	63.9	82.1	71.9	↑
	COMMONPHRASE	11.1	3.7	5.6	47.6	26.3	33.9	↑
(c)	THEMATIC	0.0	0.0	0.0	17.7	21.4	19.4	↑
	RESULTIN	0.0	0.0	0.0	50.0	33.3	40.0	↑
	HASPREREQUISITE	0.0	0.0	0.0	22.2	60.0	32.4	↑
	MATERIALMADEOF	0.0	0.0	0.0	16.7	100.0	28.6	↑
	CATEGORYEXEMPLAR	0.0	0.0	0.0	27.8	45.5	34.5	↑

Table 12: Class-wise performance of BART -EXP and BART +EXP. Relations are grouped by change in F1 after adding explanations (Δ F1): (a) relations well predicted without explanations, (b) relations can be further improved when explanations are used, (c) relations cannot be captured without context but some signals from explanations are learnt to assist the model make correct predictions.