Decomposing Generalization

Models of Generic, Habitual, and Episodic Statements

Venkata Govindarajan
University of Rochester

Benjamin Van Durme
Johns Hopkins University

Aaron Steven White
University of Rochester

Abstract

We present a novel semantic framework for modeling linguistic expressions of generalization – *generic*, *habitual*, and *episodic statements* – as combinations of simple, real-valued referential properties of predicates and their arguments. We use this framework to construct a dataset covering the entirety of the Universal Dependencies English Web Treebank. We use this dataset to probe the efficacy of type-level and token-level information – including hand-engineered features and contextual and non-contextual word embeddings – for predicting expressions of generalization.

1 Introduction

Natural language allows us to convey not only information about particular individuals and events, as in (1), but also generalizations about those individuals and events, as in (2).

(1) a. Mary ate oatmeal for breakfast today.
b. The students completed their assignments.

(2) a. Mary eats oatmeal for breakfast.
b. The students always complete their assignments on time.

This capacity for expressing generalization is extremely flexible – allowing for generalizations about the kinds of events that particular individuals are habitually involved in, as in (2), as well as characterizations about kinds of things, as in (3).

(3) a. Bishops move diagonally.
b. Soap is used to remove dirt.

Such distinctions between *episodic statements* (1), on the one hand, and *habitual* (2) and *generic (or characterizing) statements* (3), on the other, have a long history in both the linguistics and artificial intelligence literatures.[1] Nevertheless, few modern semantic parsers make a systematic distinction (though see Abzianidze and Bos 2017).

This is problematic, because the ability to accurately capture different modes of generalization is likely key to building systems with robust common sense reasoning (Zhang et al., 2017a; Bauer et al., 2018) – a central component of general artificial intelligence (McCarthy, 1960, 1980, 1986; Minsky, 1974; Schank and Abelson, 1975; Hobbs et al., 1987; Reiter, 1987). It is also surprising, since there is no dearth of data relevant to generalization (Doddington et al., 2004; Cybulska and Vossen, 2014b; Friedrich et al., 2015).

One obstacle to further progress on generalization is that current frameworks tend to take standard descriptive categories as sharp classes – e.g. EPISODIC, GENERIC, HABITUAL for statements and KIND, INDIVIDUAL for noun phrases. This may seem reasonable for sentences like (1a), where *Mary* clearly refers to a particular individual, or (3a), where *Bishops* clearly refers to a kind; but natural text is less forgiving (Grimm, 2014, 2016, 2018). Consider the underlined arguments in (4): do they refer to kinds or individuals?

(4) a. I will manage client expectations.
b. The atmosphere may not be for everyone.
c. Thanks again for great customer service!

To remedy this, we propose a novel framework for capturing linguistic expressions of generalization. Taking inspiration from *decompositional semantics* (Reisinger et al., 2015; White et al., 2016), we suggest that linguistic expressions of generalization should be captured in a continuous multi-label

[1] See Lawler 1972; Dahl 1975; Carlson 1977a,b, 1982, 1989, 1999, 2008, 2009; Geurts 1985; Declerck 1986; Schubert and Pelletier 1987, 1989; Laca 1990; Diesing 1992; Carlson and Pelletier 1995; Krifka et al. 1995; Cohen 1997, 1999, 2001, 2004; Pelletier and Asher 1997; Chierchia 1998; Kiss 1998; Prasada 2000; Schubert 2009; Van Durme 2010.

system, rather than a multi-class system. We do this by decomposing categories such as EPISODIC, HABITUAL, and GENERIC into simple referential properties of predicates and their arguments.

Using this framework (§3), we develop an annotation protocol, which we deploy (§4) to construct a new large-scale dataset of annotations covering the entire Universal Dependencies (Nivre et al., 2015) English Web Treebank (Bies et al., 2012) — the Universal Decompositional Semantics Genericity (UDS-G) dataset (available at decomp.io).

Through exploratory analysis of this dataset, we demonstrate that this multi-label framework is well-motivated (§5). We then present models for predicting expressions of linguistic generalization that combine hand-engineered type and token-level features with static and contextual learned representations (§6). We find that (i) referential properties of arguments are easier to predict than those of predicates; and that (ii) contextual learned representations contain most of the relevant information for both arguments and predicates (§7).

2 Background

Most existing annotation frameworks aim to capture expressions of linguistic generalization using multi-class annotation schemes. We argue that this reliance on multi-class annotation schemes is problematic on the basis of descriptive and theoretical work in the linguistics literature.

One of the earliest frameworks explicitly aimed at capturing expressions of linguistic generalization was developed under the **ACE-2** program (Mitchell et al., 2003; Doddington et al., 2004, and see Reiter and Frank 2010). This framework associates entity mentions with discrete labels for whether they refer to a specific member of the set in question (SPECIFIC) or any member of the set in question (GENERIC), with no formal definitions for kind- or particular-referring expressions.

The **ACE-2005** Multilingual Training Corpus (Walker et al., 2006) adds data from broadcast conversations, weblogs, and Usenet forums to the 40,106 noun phrases (NPs) from 520 newswire and broadcast documents annotated under ACE-2 and, importantly, makes changes to the genericity annotation guidelines – providing two additional classes: (i) negatively quantified entries (NEG) for referring to empty sets and (ii) underspecified entries (USP) where the referent is ambiguous between GENERIC and SPECIFIC.

The existence of the USP label already portends an issue with multi-class annotation schemes, which have no way of capturing the well-known phenomena of *taxonomic reference* (see Carlson and Pelletier, 1995, and references therein), *abstract/event reference* (Grimm, 2014, 2016, 2018), and *weak definites* (Carlson and Sussman, 2005). For example, *wines* in (5) refers to particular kinds of wine; *service* in (6) refers to an abstract entity/event that could be construed as both particular-referring, in that it is the service at a specific restaurant, and kind-referring, in that it encompasses all service events at that restaurant; and *bus* in (7) refers to potentially multiple distinct buses that are grouped into a kind by the fact that they drive a particular line.

(5) That vintner makes three different wines.

(6) The service at that restaurant is excellent.

(7) That bureaucrat takes the 90 bus to work.

A similar inflexibility is inherited by later schemes, such as **ARRAU** (Poesio et al., 2008, and see Mathew 2009; Louis and Nenkova 2011), which is mainly intended to capture anaphora resolution but which also annotates NPs for a binary GENERIC attribute following the GNOME guidelines (Poesio, 2004). This is remedied to some extent in **ECB+** (Cybulska and Vossen, 2014b,a), which is an extension of the EventCorefBank (ECB; Bejan and Harabagiu, 2010; Lee et al., 2012) – which annotates Google News texts for event coreference in accordance with the TimeML specification (Pustejovsky et al., 2003). ECB+ is an improvement in the sense that event and entity mentions may be labeled with a GENERIC class.

The ECB+ approach is useful, since episodic, habitual, and generic statements can straightforwardly be described using combinations of event and entity mention labels. For example, episodic statement will involve only non-generic entity and event mentions; habitual statements will involve a generic event mention and at least one non-generic entity mention; and generic statements will only involve generic event and entity mentions. This demonstrates the strength of decomposing statements into properties of the events and entities they describe; but there remain difficult issues arising from the fact that the decomposition does not go far enough. One is that, like ACE-2/2005 and ARRAU, ECB+ does not make it possible to capture taxonomic and abstract reference or weak definites; another is that, because ECB+ treats

generics as mutually exclusive from other event classes, it is not possible to capture that events and states in those classes can themselves be particular or generic. This is well-known for different classes of events, such as those determined by a predicate's *lexical aspect* (Vendler, 1957); but it is likely also important for distinguishing more particular *stage-level properties* – e.g. availability (8) – from more generic *individual-level properties* – e.g. strength (9) (Carlson, 1977a).

(8) Those firemen are available.

(9) Those firemen are strong.

This situation is improved upon in the Richer Event Descriptions (**RED**; O'Gorman et al., 2016) and Situation Entities (**SitEnt**; Friedrich and Palmer, 2014a,b; Friedrich et al., 2015; Friedrich and Pinkal, 2015b,a; Friedrich et al., 2016) frameworks, which annotate both NPs and entire clauses for genericity. In particular, SitEnt, which is used to annotate MASC (Ide et al., 2010) and Wikipedia, has the nice property that it recognizes the existence of abstract entities and lexical aspectual class of clauses' main verbs, along with habituality and genericity. This is useful because, in addition to decomposing statements using the genericity of the main referent and event, this framework recognizes that lexical aspect is an independent phenomenon. In practice, however, the annotations produced by this framework are mapped into a multi-class scheme containing only the high-level GENERIC-HABITUAL-EPISODIC distinction – alongside a conceptually independent distinction among illocutionary acts.

A potential argument in favor of mapping into a multi-class scheme is that, if it is sufficiently elaborated, the relevant decomposition may be recoverable. But regardless of such an elaboration, uncertainty about which which class any particular entity or event falls into cannot be ignored. Some examples may just not have categorically correct answers; and even if they do, annotator uncertainty and bias may obscure them. To account for this, we develop a novel annotation framework that both (i) explicitly captures annotator confidence about the different referential properties discussed above and (ii) automatically corrects for annotator bias using standard psycholinguistic methods.

3 Annotation Framework

We divide our framework into two protocols – the *argument* and *predicate protocols* – that probe

Figure 1: Examples of argument protocol (top) and predicate protocol (bottom) for the sentence *I will manage client expectations accordingly*.

properties of individuals and situations – i.e. events or states – referred to in a clause. A crucial aspect of our framework is that (i) multiple properties can be simultaneously true for a particular individual or situation; and (ii) we explicitly collect confidence ratings for each property. This makes our framework highly extensible, since further properties can be added without breaking a strict multi-class ontology.

We focus on properties that lie along three main axes: whether a predicate or its arguments refer to (i) instantiated or spatiotemporally delimited – i.e. *particular* situations or individuals; (ii) classes of situations – i.e. *hypothetical* situations or *kinds* of individuals; and/or (iii) intangible – i.e. *abstract* or *stative* situations or individuals.

Figure 1 shows examples of the argument protocol (top) and predicate protocol (bottom), whose implementation is based on the event factuality annotation protocol described by White et al. (2016) and Rudinger et al. (2018). Annotators are presented with a sentence with one or many words highlighted, followed by statements pertaining to the highlighted words in the context of the sentence.[2] They are then asked to fill in the statement with a binary response saying whether it *does* or *does not* hold and to give their confidence on a 5 point scale – *not at all confident* (1), *not very confident* (2), *somewhat confident* (3), *very confident* (4), and *totally confident* (5). The task instructions, along with the protocol implementation, are available at decomp.io.

[2]If the predicate head is a verb, only that verb is highlighted; if it is, copular the entire predicate is highlighted.

4 Data Collection

We use our annotation framework to collect annotations of predicates and arguments in the Universal Dependencies (Silveira et al., 2014; De Marneffe et al., 2014) English Web Treebank (Bies et al., 2012) – thus yielding the Universal Decompositional Semantics Genericity (UDS-G) dataset. UD-EWT has three main advantages over other similar corpora: (i) it contains text from multiple genres, not just newswire; and (ii) it contains gold standard Universal Dependency parses; and (iii) there are now a wide variety of other semantic annotations using the same predicate-argument extraction standard (White et al., 2016; Zhang et al., 2017b; Rudinger et al., 2018). Table 1 compares our dataset against other large annotated resources for generalization. Our data collection procedure had four stages: (i) predicate-argument extraction; (ii) predicate-argument filtering; (iii) bulk annotation; and (iv) rating normalization.

Predicate-argument extraction We extract predicates and arguments using PredPatt (White et al., 2016; Zhang et al., 2017b), which identified 34,025 predicates and 56,246 arguments of those predicates from 16,622 sentences. The parameters used for extraction are shipped with the code.

Predicate and argument filtering Based on analysis of pilot data, we developed a set of heuristics for filtering certain tokens that PredPatt identifies as predicates and arguments, either because we found that there was little variability in the label assigned to particular subsets of tokens – e.g. pronominal arguments, such as *I*, *we*, *he*, *she*, etc., are almost always labeled particular, non-kind, and non-abstract (with the exception of *you* and *they*, which can be kind-referring) – or because it is not generally possible to answer questions about those tokens – e.g. adverbial predicates are excluded. A full specification of these filtering heuristics is shipped with the data. Based on these filtering heuristics, we retain 37,146 arguments and 33, 114 predicates for annotation.

Bulk annotation 482 annotators were recruited from Amazon Mechanical Turk to annotate arguments; and 438 annotators were recruited to annotate predicates. Arguments and predicates in the UD-EWT validation and test sets were annotated by three annotators each; and those in the UD-EWT train set were annotated by one each.

Corpus	Level	Scheme	Size
ACE-2 ACE-2005	NP	multi-class	40,106
ECB+	Arg. Pred.	multi-class multi-class	12,540 14,884
CFD	NP	multi-class	3,422
Matthew et al	clause	multi-class	1,052
ARRAU	NP	multi-class	91,933
SitEnt	Topic Clause	multi-class multi-class	40,940
RED	Arg. Pred.	multi-class multi-class	10,319 8,731
UDS-G	**Arg.** **Pred.**	**multi-label** **multi-label**	**37,146** **33,114**

Table 1: Survey of genericity annotated corpora for English, including our new corpus (in bold).

Annotation normalization The need to adjust annotations biases introduced by different annotators has long been recognized in the psycholinguistics literature (Baayen, 2008) and is often carried out using mixed effects models (Gelman and Hill, 2014) and/or rating normalization procedures, such as z-scoring or ridit scoring (Agresti, 2003). We employ such procedures with the aim of producing a single real-valued score for each property that accounts for annotator confidence while adjusting for annotator bias.

Confidence normalization Different annotators use the confidence scale in different ways – e.g. some annotators use all five options while others only ever respond with *totally confident* (5). To adjust for these differences, we normalize the confidence ratings for each property using a standard ordinal scale normalization technique known as ridit scoring. In ridit scoring ordinal labels are mapped to (0, 1) using the empirical cumulative distribution function of the ratings given by each annotator. Specifically, for the responses $\mathbf{y}^{(a)}$ given by annotator a, $\text{ridit}_{\mathbf{y}^{(a)}}\left(y_i^{(a)}\right) = \text{ECDF}_{\mathbf{y}^{(a)}}\left(y_i^{(a)} - 1\right) + 0.5 \times \text{ECDF}_{\mathbf{y}^{(a)}}\left(y_i^{(a)}\right)$.

Ridit scoring has the effect of reweighting the importance of a scale label based on the frequency with which it is used. For example, insofar as an annotator rarely uses extreme values, such as *not at all confident* or *totally confident*, the annotator is likely signaling very low or very high confidence, respectively, when they are used; and insofar as an annotator often uses extreme values, the

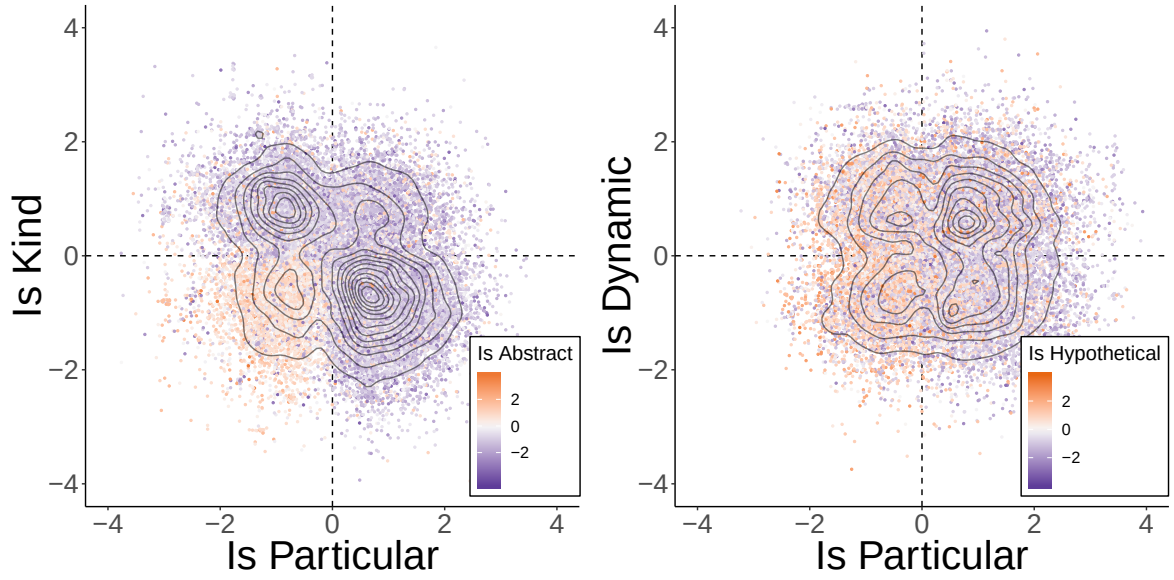

Figure 2: Distribution of normalized annotations in argument (left) and predicate (right) protocols.

annotator is likely not signaling particularly low or particularly high confidence.

Binary normalization In analyzing pilot data, we found that different annotators also have different biases for responding *true* or *false* on different properties. To adjust for these biases, we construct a normalized score using mixed effects logistic regressions fit separately to our train and development splits and our test splits. These mixed effects models all had (i) a hinge loss with margin set to the normalized confidence rating; (ii) fixed effects for property – PARTICULAR, KIND, and ABSTRACT for arguments; PARTICULAR, HYPOTHETICAL, and DYNAMIC for predicates – token, and their interaction; and (iii) by-annotator random intercepts and random slopes for property with diagonal covariance matrices. We obtain a normalized score from these models by setting the Best Linear Unbiased Predictors for the by-annotator random effects to zero and using the Best Linear Unbiased Estimators for the fixed effects to obtain a real-valued label for each token on each property. This procedure amounts to estimating a label for each property and each token based on the 'average annotator.'

5 Exploratory Analysis

Before presenting models for predicting our properties, we conduct a variety of exploratory analyses to demonstrate that the properties of the dataset relate to other token- and type-level semantic properties in intuitive ways.

Figure 2 plots the normalized ratings for the argument (left) and predicate (right) protocols. Each point corresponds to a token and the density plots visualize the number of points in a region.

Arguments We see that arguments have a clear tendency (Pearson correlation ρ=-0.33) to refer to either a kind or a particular – e.g. *place* in (10) falls in the lower right quadrant (particular-referring) and *transportation* in (11) falls in the upper left quadrant (kind-referring) – though there are a not insignificant number of arguments that refer to something that is both – e.g. *registration* in (12) falls in the upper right quadrant.

(10) I think this place is probably really great especially judging by the reviews on here .

(11) What made it perfect was that they offered transportation so that...

(12) Some places do the registration right at the hospital...

We also see that there is a clear tendency for arguments that are neither particular-referring (ρ=-0.28) nor kind-referring (ρ=-0.11) to be abstract-referring – e.g. *power* in (13) falls in the lower left quadrant (only abstract-referring) – but that there are some arguments that refer to abstract kinds and some that refer to abstract particulars – e.g. both *reputation* (14) and *argument* (15) are abstract, but *reputation* falls in the lower right quadrant, while *argument* falls in the upper left (kind-referring).

(13) Power be where power lies.

(14) Meanwhile, his reputation seems to be improving, although Bangs noted a "pretty interesting social dynamic."

(15) The Pew researchers tried to transcend the economic argument.

Predicates We see that there is effectively no tendency (ρ=0.00) for predicates that refer to particular situations to refer to dynamic events – e.g. *faxed* in the (16) falls in the upper right quadrant (particular- and dynamic-referring), while *available* in (17) falls in the lower right quadrant (particular- and non-dynamic-referring).

(16) I have faxed to you the form of Bond...

(17) is gare montparnasse storage still available?

But we do see that there is a clear tendency (ρ=-0.25) for predicates that are hypothetical-referring not to be particular-referring – e.g *knows* in (18a) and *do* in (18b) are hypotheticals in the lower left.

(18) a. Who knows what the future might hold , and it might be expensive ?

b. I have tryed to give him water but he wont take it..what should i do?

Inducing clause types One impetus for developing a multi-label framework for capturing linguistic expressions of generalization was that three-way classification of clauses into EPISODIC, GENERIC, and HABITUAL appeared insufficient. If it were sufficient, we would expect that clauses represented using our multi-label framework should cluster into three (or at least some small number of) distinct groups.

To check this, we concatenate the normalized ratings for each argument with the normalized ratings for its corresponding predicate and fit a Gaussian Mixture Model (GMM) with a Dirichlet Process (DP) prior to these predicate-argument pairs. Even with concentration parameters set to induce high sparsity ($\alpha = 0.01$), this method assigns only 22% of the predicate-argument pairs to the three most populous categories – seen in Figure 3.

Comparison to other token-level properties We compare our token-level argument and predicate properties against argument and predicate properties found in two other token-level datasets.

Event Factuality We expect that event hypotheticality should be related to event factuality – i.e. whether an event happened or not. Specifically, we expect hypothetical events to tend not to be factual. To test this, we use an event factuality dataset annotated on UD-EWT, developed by White et al. (2016) and Rudinger et al. (2018). In this dataset, all verbal predicates produced by PredPatt are annotated for whether the event they refer to already happened or is currently happening, along with a confidence rating on a five-point scale. We apply the same normalization procedure used for our properties to the factuality data and compare our normalized predicate properties against this normalized factuality score. We find that 78% of the predicates annotated in the train and dev portions of our dataset were also annotated in the factuality dataset. Among these predicates, we corroborate our expectations, finding a Spearman correlation with IS.FACTUAL of -0.25 for IS.HYPOTHETICAL, 0.12 for IS.PARTICULAR and 0.02 for IS.DYNAMIC.

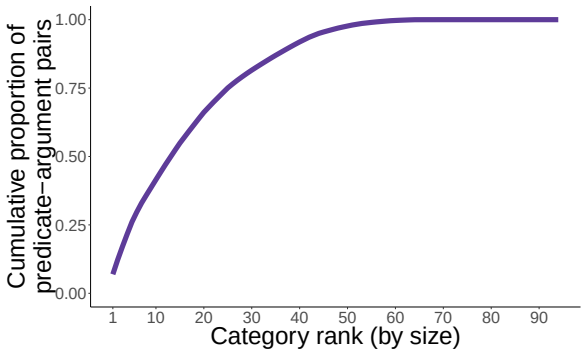

Figure 3: Cumulative distribution of predicate-argument pairs in N most frequent categories from GMM with DP prior ($\alpha = 0.01$).

Semantic Proto-Role Properties Referential properties have long been known to be important for determining argument-taking behavior in ways similar to the semantic proto-role properties of Dowty (1991) – see, e.g., the noun incorporation literature (Mithun, 1984, 1986; Baker, 1988; Van Geenhoven and Van Geenhoven, 1998; Farkas and Swart, 2003; Massam, 2009). We thus expect some amount of correlation between our properties and proto-role properties. Reisinger et al. (2015) present an annotation framework for semantic properties relevant to determining semantic role based on Dowty's (1991) seminal work. This framework was then updated and applied to UD-EWT by White et al. (2016). Table 2 gives the correlation between the ridit normalized ratings for various SPR properties on argument spans, with our argument properties. We see that properties that are associated with agentivity (AWARENESS, VOLITION, INSTIGATION, etc.) correlate positively with particularity and negatively with abstractness and (to some extent) kindhood.

Comparison to type-level properties We compare our token-level argument and predicate properties against argument and predicate properties found in two type-level datasets.

Property	Is Part	Is Kind	Is Abs
awareness	0.16	-0.1	-0.15
volition	0.16	-0.11	-0.15
sentient	0.16	-0.08	-0.16
instigation	0.10	-0.08	-0.09
existed before	0.16	-0.04	-0.17
existed during	0.10	-0.02	-0.07
existed after	0.15	-0.06	-0.14
was for benefit	0.11	-0.08	-0.11
change of location	0.07	0.06	-0.17
change of state	-0.02	0.03	-0.03
was used	0.08	-0.03	-0.09
change of possession	-0.04	0.11	-0.04
partitive	-0.02	0.04	-0.06

Table 2: Spearman correlation of Argument protocol properties with SPR properties

Eventivity The LCS Database contains hand-built lexicoconceptual structures, from which predicate eventivity and stativity can be inferred based on whether or not a particular sense contains a root node `be` (Dorr and Voss, 1993). We compare our IS.DYNAMIC predicate annotations against the eventivity ratings of verb lemmas from LCS. If a lemma possessed at least one LCS structure (sense) where it had a dynamic or stative reading, we consider it to be dynamic or stative (or both). 43.7% of the predicate lemmas in our dataset were present in the LCS database, and by thresholding the normalized scores for IS.DYNAMIC at zero – greater than 0 is dynamic, less than is not dynamic – we observe that 86.4% of predicates share at least one sense in which both are eventive or stative and 40.9% share all senses. For example, the lemmas *exist*, *thrive*, and *take* contain eventive and stative senses in both the LCS database and our annotations.

Concreteness The Concreteness rating lexicon provides concreteness ratings, which evaluate the degree to which the concept denoted by a word refers to a perceptible entity, for 40,000 generally known English lemmas (Brysbaert et al., 2014). We compare our IS.ABSTRACT argument annotations against these ratings. Concreteness ratings were found for 66% of the lemmas in our dataset, and the normalized IS.ABSTRACT score exhibited a Spearman correlation of -0.45.

6 Models

We consider two forms of predicate and argument representations to predict the three attributes in our framework: hand-engineered features and learned features. For both, we contrast both type-level information and token-level information.

Hand-engineered features We consider five sets of type-level hand-engineered features.

1. *Concreteness* Concreteness ratings for root argument lemmas in the argument protocol from the concreteness database (Brysbaert et al., 2014). For the predicate protocol, we assign 3 concreteness features: the mean, maximum and minimum concreteness rating of its arguments.
2. *Eventivity* Eventivity and stativity for the root predicate lemma in the predicate protocol and the predicate head of the root argument in the argument protocol from the LCS database.
3. *VerbNet* Verb classes from VerbNet (Schuler, 2005) for predicate lemmas.
4. *FrameNet* Frames evoked by root predicate lemmas in the predicate protocol and for both the root argument lemma and its predicate head in the argument protocol from FrameNet (Baker et al., 1998).
5. *WordNet* WordNet (Fellbaum, 1998) *supersenses* (Ciaramita and Johnson, 2003) for argument and predicate lemmas.

And we consider two sets of token-level hand-engineered features.

1. *Syntactic features* POS tags, UD morphological features, and governing dependencies were extracted using PredPatt for the predicate/argument root and all of its dependents.
2. *Lexical features* Function words – determiners, modals, auxiliaries – in the dependents of the annotated arguments and predicates.

Learned features For our type-level learned features, we use the 42B uncased GloVe embeddings for the root of the annotated predicate or argument (Pennington et al., 2014). For our token-level learned features, we use 1,024-dimensional ELMO embeddings (Peters et al., 2018). To obtain the latter, the UD-EWT sentences are passed as input to the ELMO three-layered biLM, and we extract the output of all three hidden layers for the root of the annotated predicates and arguments, giving us 3,072-dimensional vectors for each.

Labeling models For each protocol, we predict the three normalized properties corresponding to the annotated token(s) using different subsets of the above features. The feature representation is used as the input to a multilayer perceptron with ReLU nonlinearity and L1 loss. The number of hidden layers and the hidden layer sizes are hyperparameters that we tune on the development set.

Implementation For all experiments, we use stochastic gradient descent to train the multi-layer neural network parameters with the Adam optimizer (Kingma and Ba, 2014), using the default learning rate in pytorch (1e-3). We performed ablation experiments on the 4 major classes of features discussed above.

Hyperparameters For each of the ablation experiments, we ran a hyperparameter grid search over hidden layer sizes (one or two hidden layers with sizes 512, 256, 128, 64, 32; the second layer at most half the size of the first), L2 regularization penalty (0, 0.00001, 0.0001, 0.001), and the dropout probability (0.1, 0.2, 0.3, 0.4, 0.5).

Development For all models, we train for at most 20 epochs with early stopping. At the end of each epoch, the L1 loss is calculated on the development set, and if it is higher than the previous epoch, we stop training, saving the parameter values from the previous epoch.

Evaluation Consonant with work in event factuality prediction, we report Pearson correlation (ρ) and proportion of mean absolute error (MAE) explained by the model, which we refer to as R1 on analogy with the variance explained R2 = ρ^2.

$$\text{R1} = 1 - \frac{\text{MAE}^p_{\text{model}}}{\text{MAE}^p_{\text{baseline}}}$$

where $\text{MAE}^p_{\text{baseline}}$ is always guessing the median for property p. We calculate R1 across properties (wR1) by taking the mean R1 weighted by the MAE for each property.

These metrics together are useful, since ρ tells us how similar the predictions are to the true values, ignoring scale, and R1 tells us how close the predictions are to the true values, after accounting for variability in the data. We focus mainly on differences in relative performance among our models on these metrics, but for comparison, state-of-the-art event factuality prediction systems obtain $\rho \approx 0.77$ and R1 ≈ 0.57 for predicting event factuality on the predicates we annotate.

7 Results

Table 3 contains the results on the test set for both the argument (top) and predicate (bottom) protocols. We see that (i) our models are generally better able to predict referential properties of arguments than those of predicates; (ii) for both predicates and arguments, contextual learned representations contain most of the relevant information for both arguments and predicates, though the addition of hand-engineered features can give a slight performance boost, particularly for the predicate properties; and (iii) the results for proportion absolute error explained are significantly lower than what we might expect from the variance explained implied by the correlations. We discuss (i) and (ii) here, deferring discussion of (iii) to §8.

Argument properties While type-level hand-engineered and learned features perform relatively poorly for properties such as IS.PARTICULAR and IS.KIND for arguments, they are able to predict IS.ABSTRACT relatively well compared to the models with all features. The converse of this also holds: token-level hand-engineered features are better able to predict IS.PARTICULAR and IS.KIND, but perform relatively poorly on their own for IS.ABSTRACT.

This seems likely to be a product of abstract reference being fairly strongly associated with particular lexical items, while most arguments can refer particulars and kinds and which they refer to is context-dependent. And in light of the relatively good performance of contextual learned features alone, it suggests that these contextual learned features – in contrast to the hand-engineered token-level features – are able to use this information coming from the lexical item.

Interestingly, however, the models with both contextual learned features (ELMo) and hand-engineered token-level features perform slightly better than those without the hand-engineered features across the board, suggesting that there is some (small) amount of contextual information relevant to generalization that the contextual learned features are missing. This performance boost may be diminished by improved contextual encoders, such as BERT (Devlin et al., 2018).

Predicate properties We see a pattern similar to the one observed for the argument properties mirrored in the predicate properties: while type-level hand-engineered and learned features perform relatively poorly for properties such as IS.PARTICULAR and IS.HYPOTHETICAL, they are able to predict IS.DYNAMIC relatively well compared to the models with all features. The converse of this also holds: token-level hand-engineered features are better able to predict IS.PARTICULAR and IS.HYPOTHETICAL, but perform relatively poorly on their own for IS.ABSTRACT.

	Feature sets				Is.Particular		Is.Kind		Is.Abstract		All
	Type	Token	GloVe	ELMO	ρ	R1	ρ	R1	ρ	R1	wR1
ARGUMENT	+	-	-	-	42.4	7.4	30.2	4.9	51.4	11.7	8.1
	-	+	-	-	50.6	13.0	41.5	8.8	33.8	4.8	8.7
	-	-	+	-	44.8	10.5	33.4	3.9	47.1	9.9	8.2
	-	-	-	+	57.3	16.5	47.3	12.8	55.4	15.3	14.9
	+	+	-	-	55.3	14.1	46.2	11.6	52.6	13.0	12.9
	-	+	-	+	57.6	**17.2**	48.3	13.0	55.6	15.5	15.3
	+	+	-	+	57.8	16.7	47.8	13.1	**56.2**	**15.7**	15.2
	+	+	+	+	**58.0**	17.0	**48.4**	**13.5**	55.4	15.5	**15.4**
					Is.Particular		**Is.Hypothetical**		**Is.Dynamic**		
PREDICATE	+	-	-	-	14.0	0.8	13.4	0.0	32.5	5.6	2.0
	-	+	-	-	22.3	2.8	37.7	7.3	31.7	5.1	5.1
	-	-	+	-	20.3	2.4	22.4	1.5	27.5	3.6	2.5
	-	-	-	+	26.9	3.9	42.9	9.9	37.0	7.2	7.0
	-	-	+	+	26.2	3.8	42.6	10.0	37.3	7.3	7.0
	+	+	-	-	24.0	3.3	37.9	7.6	37.1	7.6	6.1
	-	+	-	+	26.9	4.0	**45.5**	**11.8**	**38.0**	**7.4**	**7.7**
	+	-	-	+	**28.2**	**4.3**	44.4	10.5	36.6	7.0	7.3
	+	+	+	+	26.1	3.5	43.8	10.4	37.3	7.3	7.0

Table 3: Correlation (ρ) and MAE explained (R1) on test split for argument (top) and predicate (bottom) protocols. Bolded numbers give the best result in the column; the models highlighted in blue are the ones analyzed in §8.

One caveat here is that, unlike for IS.ABSTRACT, type-level learned features (GloVe) alone perform quite poorly for IS.DYNAMIC, and the difference between the models with only type-level hand-engineered features and the ones with only token-level hand-engineered features is less stark for IS.DYNAMIC than for IS.ABSTRACT. This may suggest that, though IS.DYNAMIC is relatively constrained by the lexical item, it may be more contextually determined than IS.ABSTRACT. Another major difference between the argument properties and the predicate properties is that IS.PARTICULAR is much more difficult to predict than IS.HYPOTHETICAL. This contrasts with IS.PARTICULAR for arguments, which is easier to predict than IS.KIND.

8 Analysis

Figure 4 plots the true (normalized) property values for the argument (top) and predicate (bottom) protocols from the development set against the values predicted by the models highlighted in blue in Table 3. Points are colored by the part-of-speech of the argument or predicate root.

We see two overarching patterns. First, our models are generally reluctant to predict values outside the [-1, 1] range, despite the fact that there are not an insignificant number of of true values outside this range. This behavior likely contributes to the difference we saw between the ρ and R1 metrics, wherein R1 was generally worse than we would expect from ρ. This pattern is starkest for IS.PARTICULAR in the predicate protocol, where predictions are nearly all constrained to [0,1].

Second, the model appears to be heavily reliant on part-of-speech information – or some semantic information related to part-of-speech – for making predictions. This behavior can be seen in the fact that, though common noun-rooted arguments get relatively variable predictions, pronoun- and proper noun-rooted arguments are almost always predicted to be particular, non-kind, non-abstract; and though verb-rooted predicates also get relatively variable predictions, common noun-, adjective-, and proper noun-rooted, are almost always predicted to be non-dynamic.

Argument protocol Proper nouns tend to refer to particular, non-kind, non-abstract entities, but they can be kind-referring, which our models miss: *iPhone* in (20) and *Marines* in (19) were predicted to have low kind score and high particular score, while annotators label these arguments as non-particular and kind-referring.

(19) The US Marines took most of Fallujah Wednesday, but still face...

(20) I'm writing an essay...and I need to know if the iPhone was the first Smart Phone.

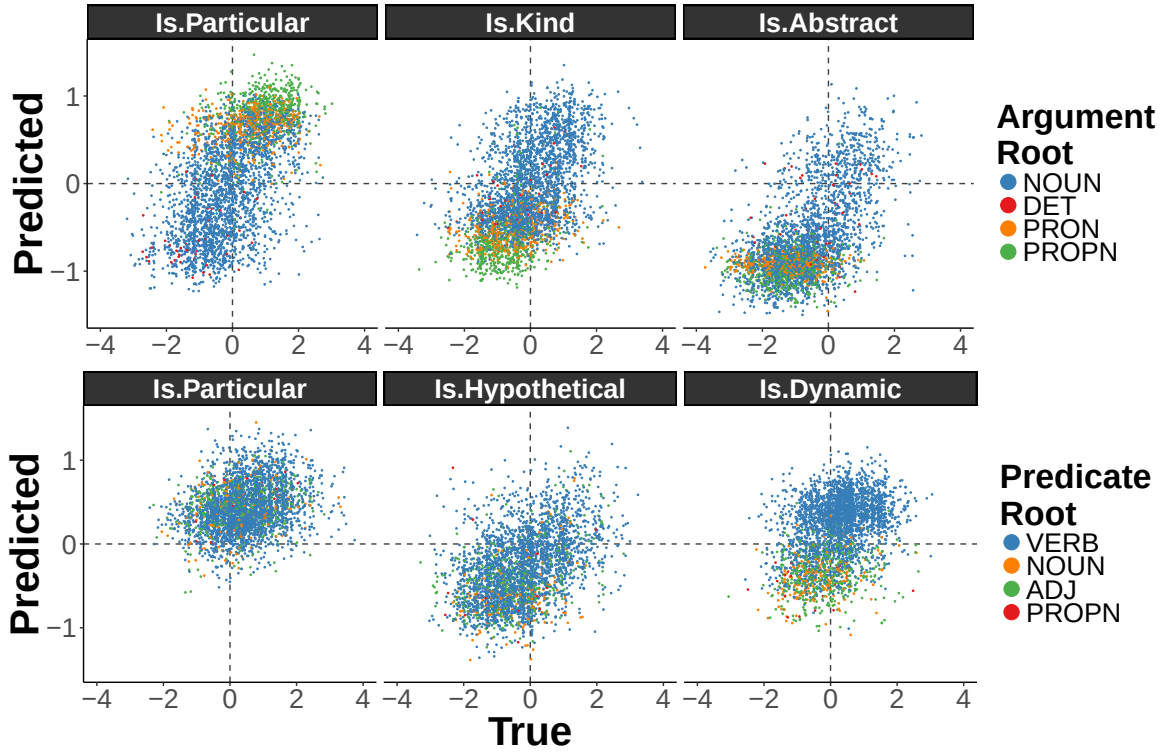

Figure 4: True (normalized) property values for argument (top) and predicate (bottom) protocols in the development set plotted against values predicted by models highlighted in blue in Table 3.

This similarly holds for pronouns. As mentioned in §4, we filtered out several pronominal arguments, but certain pronouns – like *you*, *they*, *yourself*, *themselves* – were not filtered because they can have both particular- and kind-referring uses. Our models fail to capture instances where pronouns are labeled kind-referring – e.g. *you* in (21) and (22) – consistently predicting low IS.KIND scores, likely because they are rare in our data.

(21) I like Hayes Street Grill....another plus, it's right by Civic Center, so you can take a romantic walk around the Opera House, City Hall, Symphony Auditorium...

(22) What would happen if you flew the flag of South Vietnam in Modern day Vietnam?

This behavior is not seen with common nouns: the model correctly predicts common nouns in certain contexts as non-particular, non-abstract, and kind-referring – e.g. *food* in (23) and *men* in (24).

(23) Kitchen puts out good food...

(24) just saying most men suck!

Predicate protocol As in the argument protocol, general trends associated with part-of-speech are exaggerated by the model. We noted in §5 that annotators tend to annotate hypothetical predicates as non-particular and vice-versa (ρ=-0.25), but the model's predictions are anti-correlated to a much greater extent (ρ=-0.79). For example, annotators are more willing to say a predicate can refer to particular, hypothetical situations, as in (25), or a non-particular, non-hypothetical situation, as in (26).

(25) Read the entire article; there 's a punchline...

(26) it s illegal to sell stolen property, even if you don't know its stolen.

The model also had a bias towards particular predicates referring to dynamic predicates(ρ=0.34) – a correlation not present among annotators. For instance, *is closed* in (27) was annotated as particular but non-dynamic but predicted by the model to be particular and dynamic; and *helped* in (28) was annotated as non-particular and dynamic, but the model predicted particular and dynamic.

(27) library is closed

(28) I have a new born daughter and she helped me with a lot.

9 Conclusion

We proposed a novel semantic framework for modeling linguistic expressions of generalization as combinations of simple, real-valued referential properties of predicates and their arguments. We used this framework to construct a dataset covering the entirety of the Universal Dependencies English Web Treebank.

Acknowledgments

We thank Scott Grimm and the FACTS.lab at the University of Rochester for useful comments on framework and protocol design. This research was supported by the University of Rochester, JHU HLTCOE, and DARPA AIDA. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References

Lasha Abzianidze and Johan Bos. 2017. Towards Universal Semantic Tagging. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.

Alan Agresti. 2003. *Categorical Data Analysis*, volume 482. John Wiley & Sons.

RH Baayen. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press, Cambridge.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley Framenet Project. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pages 86–90. Association for Computational Linguistics.

Mark C. Baker. 1988. *Incorporation: A theory of grammatical function changing*. University of Chicago Press Chicago.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for Generative Multi-Hop Question Answering Tasks. *arXiv preprint arXiv:1809.06309*.

Cosmin Adrian Bejan and Sanda Harabagiu. 2010. Unsupervised Event Coreference Resolution with Rich Linguistic Features. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1412–1422. Association for Computational Linguistics.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English Web Treebank. *Linguistic Data Consortium, Philadelphia, PA*.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46(3):904–911.

Greg Carlson. 2008. Patterns in the Semantics of Generic Sentences. In *Time and Modality*, pages 17–38. Springer.

Greg Carlson. 2009. Generics and Concepts. *Kinds, things, and stuff: Mass terms and generics*, pages 16–35.

Greg Carlson and Rachel Sussman. 2005. Seemingly indefinite definites. *Linguistic evidence: Empirical, theoretical, and computational perspectives*, 85:71–85.

Greg N Carlson. 1977a. *Reference to Kinds in English*. Ph.D. thesis, University of Massachusetts, Amherst.

Greg N Carlson. 1977b. A Unified Analysis of the English Bare Plural. *Linguistics and philosophy*, 1(3):413–457.

Greg N Carlson. 1982. Generic Terms and Generic Sentences. *Journal of philosophical logic*, 11(2):145–181.

Greg N Carlson. 1989. On the Semantic Composition of English Generic Sentences. In *Properties, types and meaning*, pages 167–192. Springer.

Greg N Carlson. 1999. Evaluating Generics. *Illinois Studies in the Linguistic Sciences*, 29(1):1–11.

Gregory N. Carlson and Francis Jeffry Pelletier. 1995. *The Generic Book*. University of Chicago Press. Google-Books-ID: KGBF7QvodkcC.

Gennaro Chierchia. 1998. Reference to Kinds across Language. *Natural Language Semantics*, 6(4):339–405.

Massimiliano Ciaramita and Mark Johnson. 2003. Supersense tagging of unknown nouns in WordNet. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 168–175. Association for Computational Linguistics.

A Cohen. 1997. Generics and Default Reasoning. *Computational Intelligence*, 13(4):506–533.

Ariel Cohen. 1999. Generics, frequency adverbs, and probability. *Linguistics and philosophy*, 22(3):221–253.

Ariel Cohen. 2001. On the generic use of indefinite singulars. *Journal of semantics*, 18(3):183–209.

Ariel Cohen. 2004. Generics and Mental Representations. *Linguistics and Philosophy*, 27(5):529–556.

Agata Cybulska and Piek Vossen. 2014a. Guidelines for ECB+ annotation of events and their coreference. Technical Report NWR-2014-1, VU University Amsterdam.

Agata Cybulska and Piek Vossen. 2014b. Using a sledgehammer to crack a nut? Lexical diversity and event coreference resolution. In *LREC*, pages 4545–4552.

Östen Dahl. 1975. On Generics. In Ed Keenan, editor, *Formal Semantics of Natural Language*, pages 99–111. Cambridge University Press, Cambridge.

Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of LREC*, pages 4585–4592.

Renaat Declerck. 1986. The Manifold Interpretations of Generic Sentences. *Lingua*, 68(2):149–188.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*. ArXiv: 1810.04805.

Molly Diesing. 1992. Bare Plural Subjects and the Derivation of Logical Representations. *Linguistic Inquiry*, 23(3):353–380.

George R. Doddington, Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie Strassel, and Ralph M. Weischedel. 2004. The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation. In *LREC*.

Bonnie J Dorr and Clare Voss. 1993. Machine Translation of Spatial Expressions: Defining the Relation between an Interlingua and a Knowledge Representation System. In *Proceedings of Twelfth Conference of the American Association for Artificial Intelligence, Washington, DC*, pages 374–379.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Donka F Farkas and Henriëtte de Swart. 2003. *The semantics of incorporation: From argument structure to discourse transparency*. University of Chicago Press.

C Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT press, Cambridge, MA.

Annemarie Friedrich and Alexis Palmer. 2014a. Automatic prediction of aspectual class of verbs in context. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 517–523.

Annemarie Friedrich and Alexis Palmer. 2014b. Situation Entity Annotation. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 149–158, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Annemarie Friedrich, Alexis Palmer, Melissa Peate Sørensen, and Manfred Pinkal. 2015. Annotating genericity: a survey, a scheme, and a corpus. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 21–30. Association for Computational Linguistics.

Annemarie Friedrich, Alexis Palmer, and Manfred Pinkal. 2016. Situation entity types: automatic classification of clause-level aspect. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1757–1768. Association for Computational Linguistics.

Annemarie Friedrich and Manfred Pinkal. 2015a. Automatic recognition of habituals: a three-way classification of clausal aspect. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2471–2481.

Annemarie Friedrich and Manfred Pinkal. 2015b. Discourse-sensitive automatic identification of generic expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 1272–1281.

Andrew Gelman and Jennifer Hill. 2014. *Data analysis using regression and multilevelhierarchical models*, volume 1. Cambridge University Press New York, NY, USA.

Bart Geurts. 1985. Generics. *Journal of Semantics*, 4(3):247–255.

Scott Grimm. 2014. Individuating the Abstract. In *Proceedings of Sinn und Bedeutung*, volume 18, pages 182–200.

Scott Grimm. 2016. Crime Investigations: The Countability Profile of a Delinquent Noun. *Baltic International Yearbook of Cognition, Logic and Communication*, 11(1).

Scott Grimm. 2018. Grammatical number and the scale of individuation. *Language*, 94(3):527–574.

Jerry R. Hobbs, William Croft, Todd Davies, Douglas Edwards, and Kenneth Laws. 1987. Commonsense Metaphysics and Lexical Semantics. *Comput. Linguist.*, 13(3-4):241–250.

Nancy Ide, Christiane Fellbaum, Collin Baker, and Rebecca Passonneau. 2010. The manually annotated sub-corpus: A community resource for and by the people. In *Proceedings of the ACL 2010 conference short papers*, pages 68–73. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Katalin É. Kiss. 1998. On Generic and Existential Bare Plurals and the Classification of Predicates. In Susan Rothstein, editor, *Events and Grammar*, Studies in Linguistics and Philosophy, pages 145–162. Springer Netherlands, Dordrecht.

Manfred Krifka, Francis Jeffrey Pelletier, Greg Carlson, Alice ter Meulen, Godehard Link, and Gennaro Chierchia. 1995. Genericity: An introduction. In *The Generic Book*, pages 1–124. The University of Chicago Press, Chicago.

Brenda Laca. 1990. Generic objects: Some more pieces of the puzzle. *Lingua*, 81(1):25–46.

John Lawler. 1972. Generic to a fault. In *Papers from 8th Regional Meeting of the Chicago Linguistic Society*, pages 247–258, Chicago. Chicago Linguistic Society.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint Entity and Event Coreference Resolution across Documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500. Association for Computational Linguistics.

Annie Louis and Ani Nenkova. 2011. Automatic identification of general and specific sentences by leveraging discourse annotations. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 605–613. Asian Federation of Natural Language Processing.

Diane Massam. 2009. Noun incorporation: Essentials and extensions. *Language and linguistics compass*, 3(4):1076–1096.

Thomas A Mathew. 2009. Supervised categorization of habitual versus episodic sentences. Master's thesis, Georgetown University.

John McCarthy. 1960. *Programs with common sense*. RLE and MIT computation center.

John McCarthy. 1980. Circumscription—a form of non-monotonic reasoning. *Artificial Intelligence*, 13(1-2):27–39.

John McCarthy. 1986. Applications of circumscription to formalizing common sense knowledge. *Artificial Intelligence*, 28:89–116.

Marvin Minsky. 1974. A Framework for Representing Knowledge. *MIT-AI Laboratory Memo 306*.

Alexis Mitchell, Stephanie Strassel, Mark Przybocki, JK Davis, George Doddington, Ralph Grishman, Adam Meyers, Ada Brunstein, Lisa

Ferro, and Beth Sundheim. 2003. ACE-2 version 1.0 LDC2003T11. Philadelphia: Linguistic data consortium.

Marianne Mithun. 1984. The evolution of noun incorporation. *Language*, 60(4):847–894.

Marianne Mithun. 1986. On the nature of noun incorporation. *Language*, 62(1):32–37.

Joakim Nivre, Zeljko Agic, Maria Jesus Aranzabe, Masayuki Asahara, Aitziber Atutxa, Miguel Ballesteros, John Bauer, Kepa Bengoetxea, Riyaz Ahmad Bhat, Cristina Bosco, Sam Bowman, Giuseppe G. A. Celano, Miriam Connor, Marie-Catherine de Marneffe, Arantza Diaz de Ilarraza, Kaja Dobrovoljc, Timothy Dozat, Tomaž Erjavec, Richárd Farkas, Jennifer Foster, Daniel Galbraith, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Berta Gonzales, Bruno Guillaume, Jan Hajič, Dag Haug, Radu Ion, Elena Irimia, Anders Johannsen, Hiroshi Kanayama, Jenna Kanerva, Simon Krek, Veronika Laippala, Alessandro Lenci, Nikola Ljubešić, Teresa Lynn, Christopher Manning, Cătălina Mărănduc, David Mareček, Héctor Martínez Alonso, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Anna Missilä, Verginica Mititelu, Yusuke Miyao, Simonetta Montemagni, Shunsuke Mori, Hanna Nurmi, Petya Osenova, Lilja Øvrelid, Elena Pascual, Marco Passarotti, Cenel-Augusto Perez, Slav Petrov, Jussi Piitulainen, Barbara Plank, Martin Popel, Prokopis Prokopidis, Sampo Pyysalo, Loganathan Ramasamy, Rudolf Rosa, Shadi Saleh, Sebastian Schuster, Wolfgang Seeker, Mojgan Seraji, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Kiril Simov, Aaron Smith, Jan Štěpánek, Alane Suhr, Zsolt Szántó, Takaaki Tanaka, Reut Tsarfaty, Sumire Uematsu, Larraitz Uria, Viktor Varga, Veronika Vincze, Zdeněk Žabokrtský, Daniel Zeman, and Hanzhi Zhu. 2015. Universal Dependencies 1.2. *http://universaldependencies.github.io/docs/*.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer Event Description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

Francis Jeffry Pelletier and Nicholas Asher. 1997. Generics and Defaults. In *Handbook of logic and language*, pages 1125–1177. Elsevier.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Massimo Poesio. 2004. Discourse annotation and semantic annotation in the gnome corpus. In *Proceedings of the 2004 ACL Workshop on Discourse Annotation*, pages 72–79. Association for Computational Linguistics.

Massimo Poesio, Ron Artstein, et al. 2008. Anaphoric Annotation in the ARRAU Corpus. In *LREC*.

Sandeep Prasada. 2000. Acquiring generic knowledge. *Trends in Cognitive Sciences*, 4(2):66–72.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

Drew Reisinger, Rachel Rudinger, Francis Ferraro, Craig Harman, Kyle Rawlins, and Benjamin Van Durme. 2015. Semantic Proto-Roles. *Transactions of the Association for Computational Linguistics*, 3:475–488.

Nils Reiter and Anette Frank. 2010. Identifying Generic Noun Phrases. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 40–49. Association for Computational Linguistics.

Raymond Reiter. 1987. Nonmonotonic reasoning. In J.F. Traub, N.J. Nilsson, and B.J. Grosz, editors, *Annual Review of Computer Science*, pages 147–186. Annual Reviews Inc., Palo Alto.

Rachel Rudinger, Aaron Steven White, and Benjamin Van Durme. 2018. Neural Models of Factuality. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 731–744, New Orleans, Louisiana. Association for Computational Linguistics.

Roger C. Schank and Robert P. Abelson. 1975. Scripts, Plans, and Knowledge. In *Proceedings of the 4th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'75, pages 151–157. Morgan Kaufmann Publishers Inc.

Lenhart Schubert. 2009. From generic sentences to scripts. *Logic and the Simulation of Interaction and Reasoning*, page 19.

Lenhart K Schubert and Francis Jeffry Pelletier. 1987. Problems in the representation of the logical form of generics, plurals, and mass nouns. In *New Directions in Semantics*, pages 385–451. Academic Press, London.

Lenhart K Schubert and Francis Jeffry Pelletier. 1989. Generically speaking, or, using discourse representation theory to interpret generics. In *Properties, types and meaning*, pages 193–268. Springer.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Computer and Information Science Department, Universiy of Pennsylvania.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A Gold Standard Dependency Corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Benjamin Van Durme. 2010. *Extracting Implicit Knowledge from Text*. Ph.D. thesis, University of Rochester.

Veerle Van Geenhoven and Veerle Van Geenhoven. 1998. *Semantic incorporation and indefinite descriptions: Semantic and syntactic aspects of noun incorporation in West Greenlandic*. CSLI Publications, Stanford.

Zeno Vendler. 1957. Verbs and Times. *Philosophical Review*, 66(2):143–160.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus LDC2006T06. Philadelphia: Linguistic data consortium.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1713–1723, Austin, TX. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017a. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

Sheng Zhang, Rachel Rudinger, and Benjamin Van Durme. 2017b. An Evaluation of PredPatt and Open IE via Stage 1 Semantic Role Labeling. In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.