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 static (GloVe) and contextual (ELMo) 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 Example (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 (see Carlson, 2011; Maienborn et al., 2011; Leslie and Lerner, 2016). Nevertheless, few modern semantic parsers make a systematic distinction (cf. 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): Such systems need some source for general knowledge about the world (McCarthy, 1960, 1980, 1986; Minsky, 1974; Schank and Abelson, 1975; Hobbs et al., 1987; Reiter, 1987) and natural language text seems like a prime candidate. It is also surprising, because there is no dearth of data relevant to linguistic expressions of 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—for example, 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

Transactions of the Association for Computational Linguistics, vol. 7, pp. 501–517, 2019. https://doi.org/10.1162/tacl_a_00285
Action Editor: Christopher Potts. Submission batch: 3/2019; Revision batch: 6/2019; Published 9/2019.
© 2019 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024

of generalization should be captured in a continuous multi-label 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 validate (§4) and compare against previous frameworks (§5). We then deploy this framework (§6) to construct a new large-scale dataset of annotations covering the entire Universal Dependencies (De Marneffe et al., 2014; Nivre et al., 2015) English Web Treebank (UD-EWT; Bies et al., 2012; Silveira et al., 2014)—yielding the Universal Decompositional Semantics-Genericity (UDS-G) dataset.[1] Through exploratory analysis of this dataset, we demonstrate that this multi-label framework is well-motivated (§7). 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 (§8). 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 (§9).

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). The **ACE-2005** Multilingual Training Corpus (Walker et al., 2006) extends these 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 et al., 2006). 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.

This deficit is remedied to some extent in the **ARRAU** (Poesio et al., 2018, and see Mathew, 2009; Louis and Nenkova, 2011) and **ECB+** (Cybulska and Vossen, 2014a,b) corpora. The ARRAU corpus is mainly intended to capture anaphora resolution, but following the GNOME guidelines (Poesio, 2004), it also annotates entity mentions for a GENERIC attribute, sensitive to whether the mention is in the scope of a relevant semantic operator (e.g., a conditional or quantifier) and whether the nominal refers to a type of object whose genericity is left underspecified, such as a substance. The ECB+ corpus 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), and is an improvement in the sense that, in addition to entity mentions, event 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, in ECB+, episodic statements involve only non-generic entity and event mentions; habitual statements involve a generic event mention and at least one

[1] Data, code, protocol implementation, and task instructions provided to annotators are available at decomp.io.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024

non-generic entity mention; and generic statements involve generic event mentions and at least one generic entity mention. 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, 1977).

(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 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) attempts to correct for annotator bias using standard psycholinguistic methods.

3 Annotation Framework

We divide our framework into two protocols—the *argument* and *predicate protocols*—that probe properties of individuals and situations (i.e., events or states) referred to in a clause. Drawing inspiration from prior work in *decompositional semantics* (White et al., 2016), a crucial aspect of our framework is that (i) multiple properties can be simultaneously true for a particular individual or situation (event or state); and (ii) we explicitly collect confidence ratings for each property. This makes our framework highly extensible, because further properties can be added without breaking a strict multi-class ontology.

Drawing inspiration from the prior literature on generalization discussed in §1 and §2, 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* concepts or *stative* situations).

This choice of axes is aimed at allowing our framework to capture not only the standard EPISODIC-HABITUAL-GENERIC distinction, but also phenomena that do not fit neatly into this distinction, such as taxonomic reference, abstract reference, and weak definites. The idea here is similar to prior decompositional semantics work on *semantic protoroles* (Reisinger et al., 2015; White et al., 2016, 2017), which associates categories like AGENT or PATIENT with sets of more basic properties, such as volitionality, causation, change-of-state, and so forth, and is similarly inspired by classic theoretical work (Dowty, 1991).

In our framework, prototypical episodics, habituals, and generics correspond to sets of properties that the referents of a clause's head predicate and arguments have—namely, clausal categories are built up from properties of the predicates that head

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024



Figure 1: Examples of argument protocol (top) and predicate protocol (bottom).

them along with those predicates' arguments. For instance, prototypical episodic statements, like those in (1), have arguments that only refer to particular, non-kind, non-abstract individuals and a predicate that refers to a particular event or (possibly) state; prototypical habitual statements, like those in (2) have arguments that refer to at least one particular, non-kind, non-abstract individual and a predicate that refers to a non-particular, dynamic event; and prototypical generics, like those in (3), have arguments that only refer to kinds of individuals and a predicate that refers to non-particular situations.

It is important to note that these are all *prototypical* properties of episodic, habitual, or generic statements, in the same way that volitionality is a prototypical property of agents and change-of-state is a prototypical property of patients. That is, our framework explicitly allows for bleed between categories because it only commits to the referential properties, not the categories themselves. It is this ambivalence toward sharp categories that also allows our framework to capture phenomena that fall outside the bounds of the standard three-way distinction. For instance, taxonomic reference, as in (5), and weak definites, as in (7), prototypically involve an argument being both particular- and kind-referring; stage-level properties, as in (8), prototypically involve particular, non-dynamic situations, while individual-level properties, as in (9), prototypically involve non-particular, non-dynamic situations.

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. 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).

4 Framework Validation

To demonstrate the efficacy of our framework for use in bulk annotation (reported in §6), we conduct a validation study on both our predicate and argument protocols. The aim of these studies is to establish that annotators display reasonable agreement when annotating for the properties in each protocol, relative to their reported confidence. We expect that, the more confident both annotators are in their annotation, the more likely it should be that annotators agree on those annotations.

To ensure that the findings from our validation studies generalize to the bulk annotation setting, we simulate the bulk setting as closely as possible: (i) randomly sampling arguments and predicates for annotation from the same corpus we conduct the bulk annotation on UD-EWT; and (ii) allowing annotators to do as many or as few annotations as they would like. This design makes standard measures of interannotator agreement somewhat difficult to accurately compute, because different pairs of annotators may annotate only a small number of overlapping items (arguments/predicates), so we turn to standard statistical methods from psycholinguistics to assist in estimation of interannotator agreement.

Predicate and argument extraction We extracted predicates and their arguments from the gold UD parses from UD-EWT using PredPatt (White et al., 2016; Zhang et al., 2017b). From the UD-EWT training set, we then randomly sampled 100 arguments from those headed by a `DET`, `NUM`, `NOUN`, `PROPN`, or `PRON` and 100 predicates from

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024

those headed by a ADJ, NOUN, NUM, DET, PROPN, PRON, VERB, or AUX.

Annotators A total of 44 annotators were recruited from Amazon Mechanical Turk to annotate arguments; and 50 annotators were recruited to annotate predicates. In both cases, arguments and predicates were presented in batches of 10, with each predicate and argument annotated by 10 annotators.

Confidence normalization Because 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)) we normalize the confidence ratings for each property using a standard ordinal scale normalization technique known as ridit scoring (Agresti, 2003). 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 $\mathrm{y}^{(a)}$ given by annotator a, $\text{ridit}_{\mathrm{y}^{(a)}}\left(y_i^{(a)}\right) = \text{ECDF}_{\mathrm{y}^{(a)}}\left(y_i^{(a)} - 1\right) + 0.5 \times \text{ECDF}_{\mathrm{y}^{(a)}}\left(y_i^{(a)}\right)$.

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

Interannotator Agreement (IAA) Common IAA statistics, such as Cohen's or Fleiss' κ, rely on the ability to compute both an expected agreement p_e and an observed agreement p_o, with $\kappa \equiv \frac{p_o - p_e}{1 - p_e}$. Such a computation is relatively straightforward when a small number of annotators annotate many items, but when many annotators each annotate a small number of items pairwise, p_e and p_o can be difficult to estimate accurately, especially for annotators that only annotate a few items total. Further, there is no standard way to incorporate confidence ratings, like the ones we collect, into these IAA measures.

To overcome these obstacles, we use a combination of mixed and random effects models (Gelman and Hill, 2014), which are extremely common in the analysis of psycholinguistic data (Baayen, 2008), to estimate p_e and p_o for each property. The basic idea behind using these models is to allow our estimates of p_e and p_o to be sensitive to the number of items annotators annotated as well as how annotators' confidence relates to agreement.

	Property	$\hat{\beta}_0$	$\hat{\sigma}_{\text{ann}}$	$\hat{\sigma}_{\text{item}}$
Argument	Is.Particular	0.49	1.15	1.76
	Is.Kind	−0.31	1.23	1.34
	Is.Abstract	−1.29	1.27	1.70
Predicate	Is.Particular	0.98	0.91	0.72
	Is.Dynamic	0.24	0.82	0.59
	Is.Hypothetical	−0.78	1.24	0.90

Table 1: Bias (log-odds) for answering *true*.

To estimate p_e for each property, we fit a random effects logistic regression to the binary responses for that property, with random intercepts for both annotator and item. The fixed intercept estimate $\hat{\beta}_0$ for this model is an estimate of the log-odds that the average annotator would answer *true* on that property for the average item; and the random intercepts give the distribution of actual annotator ($\hat{\sigma}_{\text{ann}}$) or item ($\hat{\sigma}_{\text{item}}$) biases. Table 1 gives the estimates for each property. We note a substantial amount of variability in the bias different annotators have for answering true on many of these properties. This variability is evidenced by the fact that $\hat{\sigma}_{\text{ann}}$ and $\hat{\sigma}_{\text{item}}$ are similar across properties, and it suggests the need to adjust for annotator biases when analyzing these data, which we do both here and for our bulk annotation.

To compute p_e from these estimates, we use a parametric bootstrap. On each replicate, we sample annotator biases b_1, b_2 independently from $\mathcal{N}(\hat{\beta}_0, \hat{\sigma}_{\text{ann}})$, then compute the expected probability of random agreement in the standard way: $\pi_1\pi_2 + (1 - \pi_1)(1 - \pi_2)$, where $\pi_i = \text{logit}_{-1}(b_i)$. We compute the mean across 9,999 such replicates to obtain p_e, shown in Table 2.

To estimate p_o for each property in a way that takes annotator confidence into account, we first compute, for each pair of annotators, each item they both annotated, and each property they annotated that item on, whether or not they agree in their annotation. We then fit separate mixed effects logistic regressions for each property to this agreement variable, with a fixed intercept β_0 and slope β_{conf} for the product of the annotators'

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024

	Property	p_e	κ_{low}	κ_{high}
Argument	Is.Particular	0.52	0.21	0.77
	Is.Kind	0.51	0.12	0.51
	Is. Abstract	0.61	0.17	0.80
Predicate	Is.Particular	0.58	−0.11	0.54
	Is.Dynamic	0.51	−0.02	0.22
	Is.Hypothetical	0.54	−0.04	0.60

Table 2: Interannotator agreement scores.

Clause type	P	R	F	κ_{mod}	κ_{ann}
EVENTIVE	0.68	0.55	0.61	0.49	0.74
STATIVE	0.61	0.59	0.60	0.47	0.67
HABITUAL	0.49	0.52	0.50	0.33	0.43
GENERIC	0.66	0.77	0.71	0.61	0.68

Table 3: Predictability of standard ontology using our property set in a kernelized support vector classifier.

confidence for that item and random intercepts for both annotator and item.[2]

We find, for all properties, that there is a reliable increase (i.e., a positive $\hat{\beta}_{\text{conf}}$) in agreement as annotators' confidence ratings go up (ps < 0.001). This corroborates our prediction that annotators should have higher agreement for things they are confident about. It also suggests the need to incorporate confidence ratings into the annotations our models are trained on, which we do in our normalization of the bulk annotation responses.

From the fixed effects, we can obtain an estimate of the probability of agreement for the average pair of annotators at each confidence level between 0 and 1. We compute two versions of κ based on such estimates: κ_{low}, which corresponds to 0 confidence for at least one annotator in a pair, and κ_{high}, which corresponds to perfect confidence for both. Table 2 shows these κ estimates.

As implied by reliably positive $\hat{\beta}_{\text{conf}}$s, we see that κ_{high} is greater than κ_{low} for all properties. Further, with the exception of DYNAMIC, κ_{high} is generally comparable to the κ estimates reported in annotations by trained annotators using a multi-class framework. For instance, compare the metrics in Table 2 to κ_{ann} in Table 3 (see §5 for details), which gives the Fleiss' κ metric for clause types in the SitEnt dataset (Friedrich et al., 2016).

5 Comparison to Standard Ontology

To demonstrate that our framework subsumes standard distinctions (e.g., EPISODIC v. HABITUAL v. GENERIC) we conduct a study comparing annotations assigned under our multi-label framework to those assigned under a framework that recognizes such multi-class distinctions. We choose the the SitEnt framework for this comparison, because it assumes a categorical distinction between GENERIC, HABITUAL (their GENERALIZING), EPISODIC (their EVENTIVE), and STATIVE clauses (Friedrich and Palmer, 2014a,b; Friedrich et al., 2015; Friedrich and Pinkal, 2015b,a; Friedrich et al., 2016).[3] SitEnt is also a useful comparison because it was constructed by highly trained annotators who had access to the entire document containing the clause being annotated, thus allowing us to assess both how much it matters that we use only very lightly trained annotators and do not provide document context.

Predicate and argument extraction For each of GENERIC, HABITUAL, STATIVE, and EVENTIVE, we randomly sample 100 clauses from SitEnt such that (i) that clause's gold annotation has that category; and (ii) all SitEnt annotators agreed on that annotation. We annotate the `mainReferent` of these clauses (as defined by SitEnt) in our argument protocol and the `mainVerb` in our predicate protocol, providing annotators only the sentence containing the clause.

Annotators 42 annotators were recruited from Amazon Mechanical Turk to annotate arguments, and 45 annotators were recruited to annotate predicates—both in batches of 10, with each predicate and argument annotated by 5 annotators.

Annotation normalization As noted in §4, different annotators use the confidence scale differently and have different biases for responding *true* or *false* on different properties (see Table 1). To adjust for these biases, we construct a normalized score for each predicate and argument using mixed effects logistic regressions. These mixed effects models all had (i) a hinge loss with margin set to the normalized confidence rating; (ii) fixed effects for property (PARTICULAR,

[2] We use the product of annotator confidences because it is large when both annotators have high confidence and small when either annotator has low confidence and always remains between 0 (lowest confidence) and 1 (highest confidence).

[3] SitEnt additionally assumes three other classes, contrasting with the four above: IMPERATIVE, QUESTION, and REPORT. We ignore clauses labeled with these categories.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024

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. The rationale behind (i) is that *true* should be associated with positive values; *false* should be associated with negative values; and the confidence rating should control how far from zero the normalized rating is, adjusting for the biases of annotators that responded to a particular item. The resulting response scale is analogous to current approaches to event factuality annotation (Lee et al., 2015; Stanovsky et al., 2017; Rudinger et al., 2018).

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.''

Quantitative comparison To compare our annotations to the gold situation entity types from SitEnt, we train a support vector classifier with a radial basis function kernel to predict the situation entity type of each clause on the basis of the normalized argument property annotations for that clause's `mainReferent` and the normalized predicate property annotations for that clause's `mainVerb`. The hyperparameters for this support vector classifier were selected using exhaustive grid search over the regularization parameter $\lambda \in \{1, 10, 100, 1000\}$ and bandwidth $\sigma \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ in a 5-fold cross-validation (CV). This 5-fold CV was nested within a 10-fold CV, from which we calculate metrics.

Table 3 reports the precision, recall, and F-score computed using the held-out set in each fold of the 10-fold CV. For purposes of comparison, it also gives the Fleiss' κ reported by Friedrich et al. (2016) for each property (κ_{ann}) as well as Cohen's κ between our model predictions on the held-out folds and the gold SitEnt annotations (κ_{mod}). One way to think about κ_{mod} is that it tells us what agreement we would expect if we used our model as an annotator instead of highly trained humans.

We see that our model's agreement (κ_{mod}) tracks interannotator agreement (κ_{ann}) surprisingly well. Indeed, in some cases, such as for GENERIC, our model's agreement is within a few points of

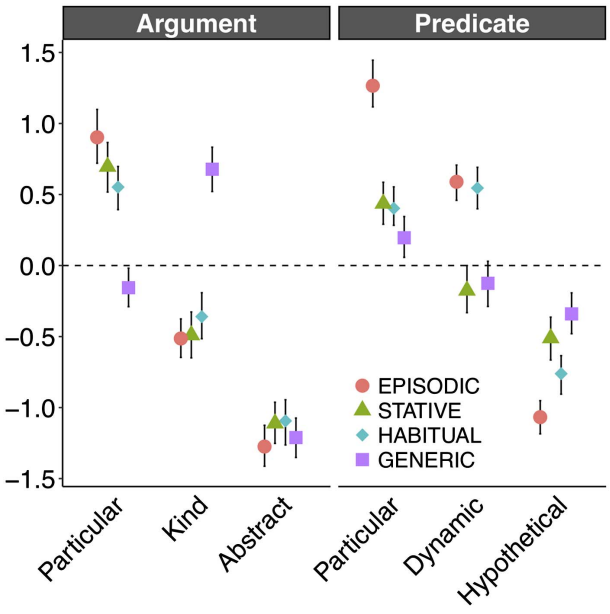


Figure 2: Mean property value for each clause type.

interannotator agreement. This pattern is surprising, because our model is based on annotations by very lightly trained annotators who have access to very limited context compared with the annotators of SitEnt, who receive the entire document in which a clause is found. Indeed, our model has access to even less context than it could otherwise have on the basis of our framework, since we only annotate one of the potentially many arguments occurring in a clause; thus, the metrics in Table 3 are likely somewhat conservative. This pattern may further suggest that, although having extra context for annotating complex semantic phenomena is always preferable, we still capture useful information by annotating only isolated sentences.

Qualitative comparison Figure 2 shows the mean normalized value for each property in our framework broken out by clause type. As expected, we see that episodics tend to have particular-referring arguments and predicates, whereas generics tend to have kind-referring arguments and non-particular predicates. Also as expected, episodics and habituals tend to refer to situations that are more dynamic than statives and generics. But although it makes sense that generics would be, on average, near zero for dynamicity—since generics can be about both dynamic and non-dynamic situations—it is less clear why statives are not more negative. This pattern may arise in some way from the fact that

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024

there is relatively lower agreement on dynamicity, as noted in §4.

6 Bulk Annotation

We use our annotation framework to collect annotations of predicates and arguments on UD-EWT using the PredPatt system—thus yielding the Universal Decompositional Semantics–Genericity (UDS-G) dataset. Using UD-EWT in conjunction with PredPatt has two main advantages over other similar corpora: (i) UD-EWT contains text from multiple genres—not just newswire—with gold standard Universal Dependency parses; and (ii) there are now a wide variety of other semantic annotations on top of UD-EWT that use the PredPatt standard (White et al., 2016; Rudinger et al., 2018; Vashishtha et al., 2019).

Predicate and argument extraction PredPatt identifies 34,025 predicates and 56,246 arguments of those predicates from 16,622 sentences. Based on analysis of the data from our validation study (§4) and other pilot experiments (not reported here), 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—for example, 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). Based on these filtering heuristics, we retain 37,146 arguments and 33,114 predicates for annotation. Table 4 compares these numbers against the resources described in §2.

Annotators We recruited 482 annotators 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. All annotations were performed in batches of 10.

Annotation normalization We use the annotation normalization procedure described in §5, fit separately to our train and development splits, on the one hand, and our test split, on the other.

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

Table 4: Survey of genericity annotated corpora for English, including our new corpus (in **bold**).

7 Exploratory Analysis

Before presenting models for predicting our properties, we conduct an exploratory analysis to demonstrate that the properties of the dataset relate to other token- and type-level semantic properties in intuitive ways. Figure 3 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 slight tendency (Pearson correlation $\rho = -0.33$) to refer to either a kind or a particular—for example, *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—for example, *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 slight tendency for arguments that are neither particular-referring ($\rho = -0.28$) nor kind-referring ($\rho = -0.11$) to be abstract-referring—for example, *power* in (13)

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024

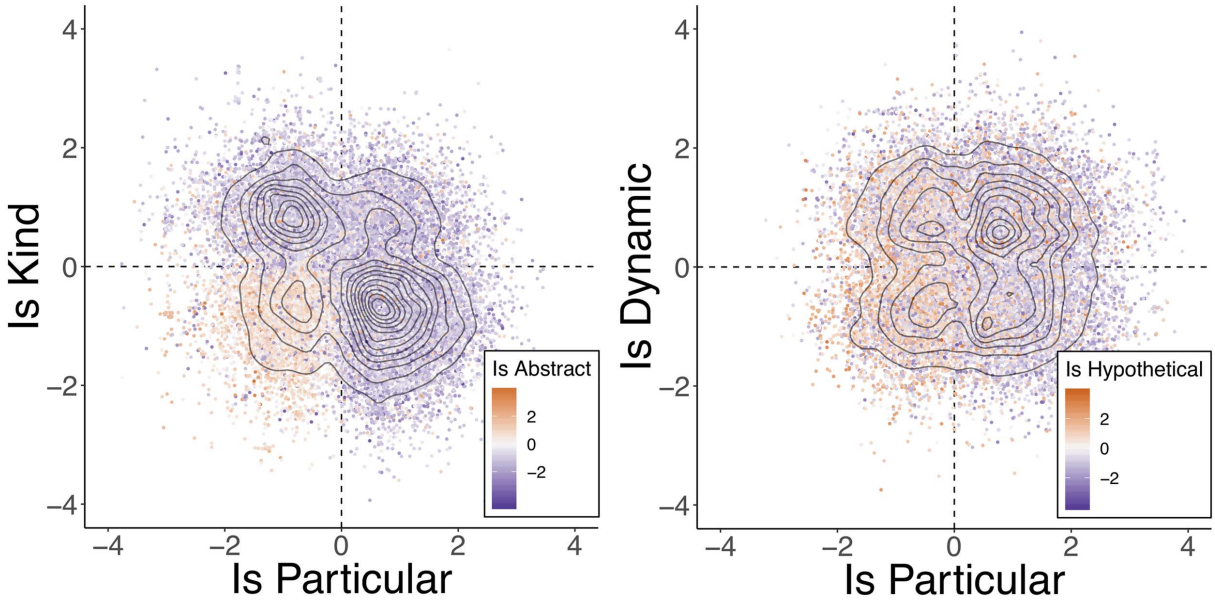


Figure 3: Distribution of normalized annotations in argument (left) and predicate (right) protocols.

falls in the lower left quadrant (only abstract-referring)—but that there are some arguments that refer to abstract particulars and some that refer to abstract kinds—for example, both *reputation* (14) and *argument* (15) are abstract, but *reputation* falls in the lower right quadrant, while *argument* falls in the upper left.

(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 ($\rho = 0.00$) for predicates that refer to particular situations to refer to dynamic events—for example, *faxed* in (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 slight tendency ($\rho = -0.25$) for predicates that are hypothetical-referring not to be particular-referring—for example, *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?

8 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) and the mean, maximum, and minimum concreteness rating of a predicate’s arguments in the predicate protocol.

2. *Eventivity* Eventivity and stativity ratings 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 (Dorr, 2001).

3. *VerbNet* Verb classes from VerbNet (Schuler, 2005) for root predicate lemmas.

4. *FrameNet* Frames evoked by root predicate lemmas in the predicate protocol and for both

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024

the root argument lemma and its predicate head in the argument protocol from FrameNet (Baker et al., 1998).

5. *WordNet* The union of WordNet (Fellbaum, 1998) *supersenses* (Ciaramita and Johnson, 2003) for all WordNet senses the root argument or predicate lemmas can have.

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 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 their sizes are hyperparameters that we tune on the development set.

Implementation For all experiments, we use stochastic gradient descent to train the multilayer perceptron parameters with the Adam optimizer (Kingma and Ba, 2015), using the default learning rate in pytorch (10^{-3}). We performed ablation experiments on the four 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 $h_1, h_2 \in \{512, 256, 128, 64, 32\}$; h_2 at most half of h_1), L2 regularization penalty $l \in \{0, 10^{-5}, 10^{-4}, 10^{-3}\}$, and the dropout probability $d \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$.

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, because 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, 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 (Rudinger et al., 2018).

9 Results

Table 5 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 proportion of absolute error explained is 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 §10.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024

	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
	−	−	+	−	44.5	8.3	33.4	4.6	45.2	7.7	6.9
	−	−	−	+	57.5	17.0	48.1	13.3	55.7	14.9	15.1
	+	+	−	−	55.3	14.1	46.2	11.6	52.6	13.0	12.9
	−	+	−	+	**58.6**	15.6	48.6	**13.7**	**56.8**	14.2	14.5
	+	+	−	+	58.3	16.3	47.8	13.2	56.3	**15.2**	14.9
	+	+	+	+	58.1	**17.0**	**48.9**	13.2	56.1	15.1	**15.1**
					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
	−	−	+	−	20.6	2.2	23.4	2.4	29.7	4.6	3.0
	−	−	−	+	26.2	3.6	43.1	10.0	37.0	6.8	6.8
	−	−	+	+	26.8	4.0	42.8	8.9	37.3	7.3	6.7
	+	+	−	−	24.0	3.3	37.9	7.6	37.1	7.6	6.1
	−	+	−	+	**27.4**	4.1	43.3	10.1	**38.6**	**7.8**	**7.4**
	+	−	−	+	27.1	4.0	43.0	10.1	37.5	7.6	7.2
	+	+	+	+	26.8	**4.1**	**43.5**	**10.3**	37.1	7.2	7.2

Table 5: 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 §10.

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 to 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., 2019).

Predicate properties We see a pattern similar to the one observed for the argument properties mirrored in the predicate properties: Whereas 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 with 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.DYNAMIC.

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

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024

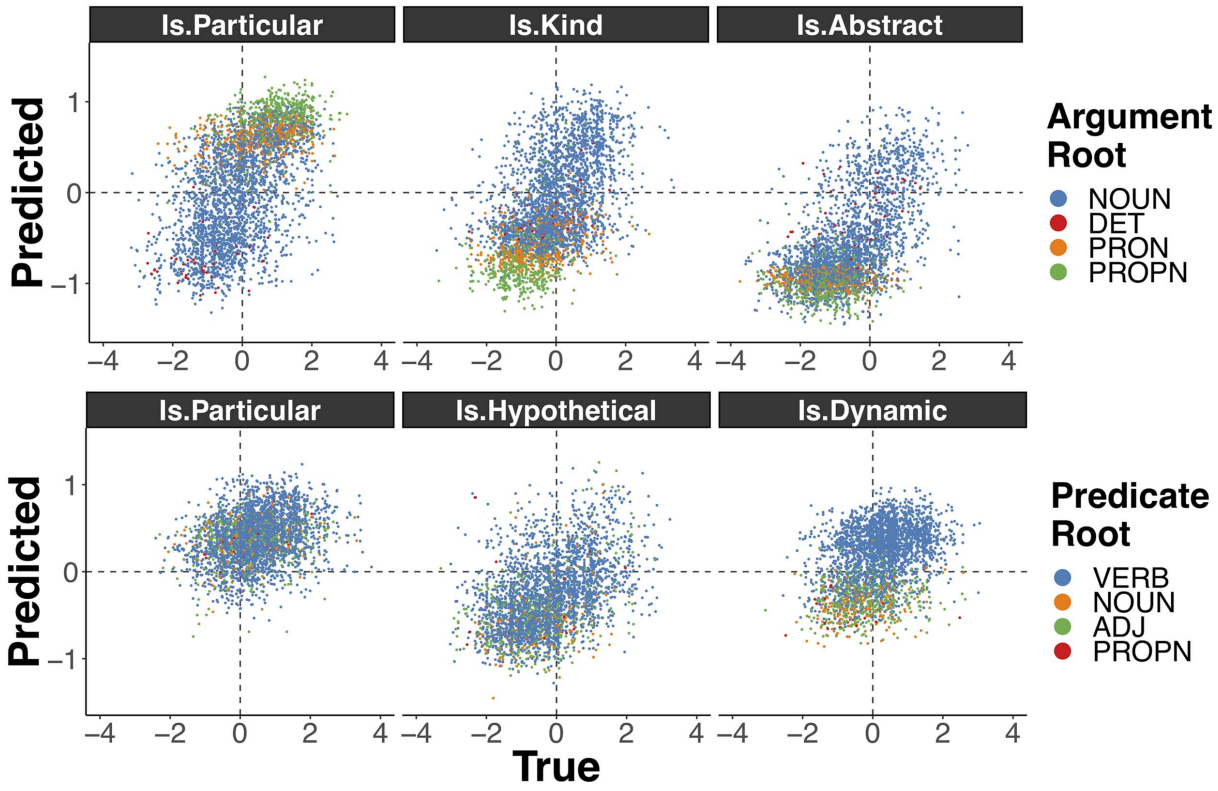


Figure 4: True (normalized) property values for argument (top) and predicate (bottom) protocols in the development set plotted against values predicted by models highlighted in blue in Table 5.

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.

10 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 5. 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 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 predicates 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.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024

This similarly holds for pronouns. As mentioned in §6, 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 §7 that annotators tend to annotate hypothetical predicates as non-particular and vice-versa ($\rho = -0.25$), but the model's predictions are anti-correlated to a much greater extent ($\rho = -0.79$). For example, annotators are more willing to say a predicate can refer to particular, hypothetical situations (25) or a non-particular, non-hypothetical situation (26).

(25) Read the entire article[...]

(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 ($\rho = 0.34$)—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.

11 Conclusion

We have 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 and probed the ability of both hand-engineered and learned type- and token-level features to predict the annotations in this dataset.

Acknowledgments

We would like to thank three anonymous reviewers and Chris Potts for useful comments on this paper as well as Scott Grimm and the FACTS.lab at the University of Rochester for useful comments on the 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*, 482. John Wiley & Sons.

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

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, pages 86–90, Montreal.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings*

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024

of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4220–4230, Brussels.

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, Uppsala.

Ann Bies, Justin Mott, Colin Warner, and Seth Kulick. 2012. English Web Treebank LDC2012T13. 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, Rachel Sussman, Natalie Klein, and Michael Tanenhaus. 2006. Weak definite noun phrases. In *Proceedings of NELS 36*, pages 179–196, Amherst, MA.

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

Gregory Carlson. 2011. Genericity, In (Maienborn et al., 2011), 1153–1185.

Gregory N. Carlson and Francis Jeffry Pelletier. 1995. *The Generic Book*, The University of Chicago Press.

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*, pages 168–175.

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 *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik.

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 the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik.

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, MN.

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 *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon.

Bonnie J. Dorr. 2001. LCS Database. University of Maryland.

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

Christiane 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)*, pages 517–523, Baltimore, MD.

Annemarie Friedrich and Alexis Palmer. 2014b. Situation entity annotation. In *Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop*, pages 149–158, Dublin.

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, Denver, CO.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024

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, Berlin.

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, Lisbon.

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)*, pages 1272–1281, Beijing.

Andrew Gelman and Jennifer Hill. 2014. *Data Analysis using Regression and Multilevel-Hierarchical Models*. Cambridge University Press, New York City.

Scott Grimm. 2014. Individuating the abstract. In *Proceedings of Sinn und Bedeutung 18*, pages 182–200, Bayonne and Vitoria-Gasteiz.

Scott Grimm. 2016. Crime investigations: The countability profile of a delinquent noun. *Baltic International Yearbook of Cognition, Logic and Communication*, 11. doi:10.4148/1944-3676.1111.

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. *Computational Linguistics*, 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, Uppsala.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of 3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, CA.

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, Jeju Island.

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer. 2015. Event detection and factuality assessment with non-expert supervision. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1643–1648, Lisbon.

Sarah-Jane Leslie and Adam Lerner. 2016. Generic generalizations, Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Winter 2016 edition. Metaphysics Research Lab, Stanford University.

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, Chiang Mai.

Claudia Maienborn, Klaus von Heusinger, and Paul Portner, editors. 2011. *Semantics: An International Handbook of Natural Language Meaning*, volume 2. Mouton de Gruyter, Berlin.

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 nonmonotonic reasoning. *Artificial Intelligence*, 13(1–2):27–39.

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

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024

Marvin Minsky. 1974. A framework for representing knowledge. MIT-AI Laboratory Memo 306.

Alexis Mitchell, Stephanie Strassel, Mark Przybocki, J. K. Davis, George Doddington, Ralph Grishman, Adam Meyers, Ada Brunstein, Lisa Ferro, and Beth Sundheim. 2003. ACE-2 Version 1.0 LDC2003T11. Linguistic Data Consortium, Philadelphia, PA.

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, Hanzhi Zhu. 2015. Universal Dependencies 1.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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, TX.

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.

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, LA.

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, Barcelona.

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22, New Orleans, LA.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The TIMEBANK corpus. In *Proceedings of Corpus Linguistics*, pages 647–656, Lancaster.

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*, pages 40–49, Uppsala.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024

Raymond Reiter. 1987. Nonmonotonic reasoning, In J. F. Traub, N. J. Nilsson, and B. J. Grosz, editors, *Annual Review of Computer Science*, volume 2, pages 147–186. Annual Reviews Inc.

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, LA.

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*, pages 151–157.

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'14)*.

Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov, Ido Dagan, and Iryna Gurevych. 2017. Integrating deep linguistic features in factuality prediction over unified datasets. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 352–357, Vancouver.

Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. *arXiv*, cs.CL/ 1902.01390v2.

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. Linguistic Data Consortium, Philadelphia, PA.

Aaron Steven White, Kyle Rawlins, and Benjamin Van Durme. 2017. The semantic proto-role linking model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2, pages 92–98, Valencia.

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.

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*, Montpellier.

Downloaded from http://www.mitpressjournals.org/doi/pdf/10.1162/tacl_a_00285 by guest on 29 March 2024