Contrast Sets for Stativity of English Verbs in Context

Daniel Chen & Alexis Palmer
Department of Linguistics
University of Colorado
{daniel.chen-1, alexis.palmer}@colorado.edu

Abstract

For the task of classifying verbs in context as dynamic or stative, current models approach human performance, but only for particular data sets. To better understand the performance of such models, and how well they are able to generalize beyond particular test sets, we apply the contrast set (Gardner et al., 2020) methodology to stativity classification. We create nearly 300 contrastive pairs by perturbing test set instances just enough to change their labels from one class to the other, while preserving coherence, meaning, and well-formedness. Contrastive evaluation shows that a model with near-human performance on an in-distribution test set degrades substantially when applied to transformed examples, showing that the stative vs. dynamic classification task is more complex than the model performance might otherwise suggest. Code and data are freely available.1

1 Introducing stativity

Aspectual properties of verbs, and the clauses they inhabit, have the potential to support a range of natural language processing tasks, such as event ordering (Modi and Titov, 2014) and temporal relation classification (Costa and Branco, 2012), as well as contributing to speaker choices around situational construal (Trott et al., 2020). At the same time, verb and situational aspect are a complex set of interacting properties, in which the meaning of the verb, the nature of its arguments, adverbial modifiers, and grammatical features such as verb tense and nominal definiteness can all play a role in determining the aspectual make-up of a clause.

This sensitivity of aspectual categorization to small shifts in linguistic form is one reason that automatic prediction of aspectual classes is an especially challenging computational problem. In this paper we explore the stability of automatic classification for one particular facet of aspect: stativity of English verbs. Stativity reflects the degree to which a verb represents a static situation versus a situation that reflects some degree of dynamicity. Dynamic verbs typically involve some change of state. (1-3) below show examples of the three classes relevant for our study: DYNAMIC verbs, STATIVE verbs, and verbs for which annotators CANNOT_DECIDE.

(1) Table 7 shows results from the latest experiments. STATIVE
(2) Dr. Smith showed her students how to work with the new GPUs. DYNAMIC
(3) The earlier paper shows the effectiveness of incorporating linguistic features. CANNOT_DECIDE

In (1), the table is static, and the results exist in the table; no change of state is indicated. (2) highlights the dynamic sense of show, in which the Agent is giving a demonstration. (3) allows two readings. In the STATIVE reading, the result about linguistic features is a static property of the paper; it simply exists in the paper. In the DYNAMIC reading, the paper demonstrates the effectiveness of linguistic features through an argument that develops and progresses over the course of the paper.

Most verbs in English have a strong predominant category (stative or dynamic), yet allow for variable interpretation, depending on context. A smaller number of verbs (e.g. show), are highly flexible, with no strong statistical tendency in either direction (Friedrich and Palmer, 2014a; Falk and Martin, 2016). Because of this variability, automatic classification of aspectual properties requires contextual input and instance-level classification.

To better understand the ability of systems to automatically determine stativity, we produce contrast sets (section 2) for English verb stativity with 298 transformed instances. The contrast set instances and their ground-truth labels are extracted from the SitEnt corpus (section 2.2), and the transformed instances are produced using a range of

1https://github.com/dchensta/se_contrast
linguistically-motivated transformation strategies, detailed in section 3.

Using a standard modeling configuration (section 4), we show that classification performance on the transformed instances is substantially lower than on the original instances. According to Friedrich (2017), observed annotator agreement for this task ranges from 79% to 82%. On the original instances, the model achieves micro-averaged accuracy of nearly 80%, approaching human agreement for this task. On the transformed instances, micro-averaged accuracy is well below 60%.

2 Building contrast sets for stativity

For many NLP tasks, neural models, especially those built on large language models, have been shown to be sensitive to annotation artifacts in the data on which they are trained and evaluated. High performance of classifiers often hinges on preserving these properties at evaluation time, and testing on out-of-distribution data can result in such dramatic performance decreases that the models no longer reliably perform the task they have been trained to do (Gururangan et al., 2018; Poliak et al., 2018; Geva et al., 2019, among others). These findings have given rise to methodologies for more careful evaluation of classification capability. Several different methods for improved evaluation have been proposed (Ribeiro et al., 2020; Gardner et al., 2020, among others).

2.1 Contrast sets

In this work, we follow the contrast set methodology (Gardner et al., 2020). The core idea is to create contrastive evaluation data sets by having experts make small perturbations to instances in the original test sets. In our case, we vary the lexical aspect of the main verb so that the preferred label changes from DYNAMIC (4) to STATIVE (5):

(4) Mary ran the Buenos Aires Marathon.

(5) Mary was a participant in the Buenos Aires Marathon.

These perturbations need to strike a delicate balance. The changes should be large enough to change the gold label for the instance, yet small enough to retain meaning, coherence, and validity. We also aim to use a variety of strategies, so as not to introduce new unintended annotation artifacts.

Once contrast sets have been built, we compare the performance of the model in question on the transformed test instances (with their new labels) to the performance of the same model on the original version of those same instances. Significant performance degradation on the transformed test data calls into question whether the model has learned to classify the phenomena modeled in the annotated training data. The contrast set consists of the paired original and transformed test instances.

2.2 Data

We use data from the SitEnt (situation entities) corpus (Friedrich and Palmer, 2014b; Friedrich, 2017). The corpus combines data from MASC (Ide et al., 2008) with Wikipedia texts, creating a collection of documents from 13 different genres. Texts are segmented into clauses, and each clause is triply-annotated for stativity of the main verb, genericity of the main referent, habituality of the situation described, and finally, a clause-level situation type label (following Smith (2003), these labels distinguish between events, states, generics, generalizing sentences, facts, propositions, reports, questions, imperatives, and undecided). Gold labels come from a majority vote across the three annotators.

For model training, we use Friedrich et al. (2016)’s original training split, which consists of 324 documents, with 42,309 clauses. As a basis for building contrast sets, we select four documents from the original test set, each from a different genre: news, essay, journal, and Wikipedia. For each sentence, we create one contrast set by transforming the first clause in the sentence.

4No contrast sets are created for clauses labeled CANNOT DECIDE.

We also aim to use a variety of strategies, so as not to introduce new unintended annotation artifacts.

\[
\begin{array}{ccc}
\text{Training} & \text{DYN} & 18,357 \\
\text{Test} & \text{STAT} & 15,507 \\
\text{Contrast:} & \text{CD} & 8,445 \\
\text{Test Orig} & 376 & 217 \\
\text{Contrast:} & 172 & 120 \\
\text{Test Trans} & 120 & 172 \\
\end{array}
\]

Table 1: Distribution of DYNAMIC (DYN), STATIVE (STAT), and CANNOT_DECIDE (CD) labels.

For model training, we use Friedrich et al. (2016)’s original training split, which consists of 324 documents, with 42,309 clauses. As a basis for building contrast sets, we select four documents from the original test set, each from a different genre: news, essay, journal, and Wikipedia. For each sentence, we create one contrast set by transforming the first clause in the sentence.

\[\text{https://github.com/annefried/sitent/tree/master/annotated_corpus}\]

\[\text{To test the viability of using first clauses only, we created contrast sets for all clauses in the Wikipedia document and compared classifier performance for the full document vs. only initial clauses. There was no significant difference.}\]
NOT_DECIDE. We produce 292 contrast sets.5

Table 1 shows the distribution of the three labels for our data set. Test refers to the four selected documents; Contrast: Test_Orig refers to the original versions for the 292 contrast sets; and Contrast: Test_Trans refers to the transformed counterparts of the original instances, with flipped labels. Note that DYNAMIC to STATIVE transformations outnum-

ber STATIVE to DYNAMIC transformations.

3 Transformation strategies

After building the contrast sets, we perform an analysis of the linguistic properties of the various strategies used. Most transformations hinge on the main verb, either replacing the lexical item or changing the role of the verb so that it moves to a different structural and semantic configuration.

3.1 DYNAMIC → STATIVE

Example sentences showing the DYNAMIC to STA-

TIVE transformations can be found in Table 2.

1. THOUGHT VERB - Demote a dynamic verb from main to secondary verb by moving it into the subordinate THEME role for verbs of thinking, believing, or feeling.

2. COPULA - Replace main verb with a simple predication headed by a copular verb.

3. DESCRIPTIVE VERB - Replace the dynamic action with a descriptive verb, effectively reconfiguring the dynamic action as stative properties of the subject noun.

4. LIGHT VERB - Use the possessive 6 light verb construction with have to make have the new main verb.

5. SEMI-MODAL - Use a semi-modal verb (e.g. need to, ought to) marking deontic modality, which concerns the speaker’s requirements and desires, as a “thought” or “emotion” from the speaker, who can be an unspecified authority with no referent.

6. DOWNGRADE TO PPL - Remove main verb from the clause by transforming it into a perfect passive participle that favors a descriptive, adjectival reading over a verbal reading.

7. ORDER - Switch the order of the clauses and insert a descriptive verb as the new main verb.

3.2 STATIVE → DYNAMIC

Example sentences showing each of the STATIVE to DYNAMIC transformations can be found in Table 3.

1. NEW PARTICIPANT - Choose a synonymous verb that introduces an agent who participates in a dynamic synonym of the original verb.

2. INSERT VERB - Replace a stative verb (typically the copula) with a dynamic verb or insert a dynamic verb as the new main verb, of which the original stative verb is a dependent.

3. BECOMING - Replace standard copula with an inflected form of the verbs to become or to get, and their synonyms. This preserves the copula’s predicating structure while reformulating the event as dynamic.

4. UPGRADE - Upgrade a perfect passive participle or subordinate STATIVE verb to the main verb of the clause by adding a helping verb or deleting the main STATIVE verb.

5. HEAVY VERB - Replace a light verb construction like have with a heavy, dynamic verb.

4 Model, results, and discussion

Having built contrast sets, we now evaluate performance compared to the original instances.

4.1 Model

Our straightforward classification model first learns a contextualized representation for each clause using BERT (Devlin et al., 2019), followed by a regression layer to classify the clause representations as DYNAMIC, STATIVE, or CD. The logistic regression model is trained using the liblinear solver and L2 regularization. Running 5-fold cross-validation over the full training set, using the trained logistic regression model, yielded accuracy scores ranging from 77.7% to 80.13%, only slightly below observed human agreement, which Friedrich (2017) reports as ranging from 79% to 82%.

4.2 Results: Classifying contrast sets

Table 4 shows the model’s performance on the original test set instances and the transformed instances. These figures include all clauses for the Wikipedia text and only initial clauses for the other three texts, and only clauses whose original labels are either STATIVE or DYNAMIC. On the original instances, the model achieves micro-averaged accuracy of nearly 80%, approaching human agreement for this task. On the transformed instances, micro-averaged accuracy is well below 60%.

5 This number is on par with the data sets described in Gardner et al. (2020), which range from 70 to 1000 contrast sets per task.

6 verbs of possession have a STATIVE reading
Since the 1960s, blackjack has been a high-profile target of advantage players, particularly card counters. During that time, the panel said, "Your actions and failure to act constituted direct violations of Senate rules." Table 2: Examples of DYNAMIC -> STATIVE transformations, along with the number of times each transformation strategy was used in the contrast sets.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Original Instance (ST)</th>
<th>Transformed Instance (DYN)</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>THOUGHT VERB</td>
<td>During that time, the panel said.</td>
<td>During that time, the panel believed that</td>
<td>55</td>
</tr>
<tr>
<td>COPULA</td>
<td>Although it affected Youngstown and the surrounding area more than it affected other regions,</td>
<td>Although it was in Youngstown and the surrounding area more than in other regions,</td>
<td>48</td>
</tr>
<tr>
<td>DESCRIPTIVE VERB</td>
<td>&quot;Your actions and failure to act led to violations of Senate rules&quot;</td>
<td>&quot;Your actions and failure to act constituted direct violations of Senate rules&quot;</td>
<td>47</td>
</tr>
<tr>
<td>LIGHT VERB</td>
<td>Scoring higher than 21</td>
<td>Having a score higher than 21</td>
<td>14</td>
</tr>
<tr>
<td>SEMI-MODAL</td>
<td>The players' initial cards may be dealt face up</td>
<td>The players' initial cards need to consistently be dealt either face up</td>
<td>5</td>
</tr>
<tr>
<td>DOWNGRADE TO PPL</td>
<td>When examining the areas history, culture, and economic situation</td>
<td>Based on the area's history, culture, and economic situation</td>
<td>2</td>
</tr>
<tr>
<td>ORDER</td>
<td>the player or the dealer wins by having a score of 21 or by having the highest score</td>
<td>having a score of 21 means the player or the dealer wins</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3: Examples of STATIVE -> DYNAMIC transformations, along with the number of times each transformation strategy was used in the contrast sets.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Original Instance (DYN)</th>
<th>Transformed Instance (ST)</th>
<th>#</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW PARTICIPANT</td>
<td>11 plus the value of any other card will always be less than or equal to 21.</td>
<td>Players add 11 plus the value of any other card to get less than or equal to 21.</td>
<td>81</td>
</tr>
<tr>
<td>INSERT VERB</td>
<td>Since the 1960s, blackjack has been a high-profile target of advantage players, particularly card counters.</td>
<td>Since the 1960s, blackjack has functioned as a high-profile target of advantage players, particularly card counters.</td>
<td>19</td>
</tr>
<tr>
<td>BECOMING</td>
<td>One such bonus was a ten-to-one payout</td>
<td>One such bonus became a ten-to-one payout</td>
<td>13</td>
</tr>
<tr>
<td>UPGRADE PPL</td>
<td>Other casino games inspired by blackjack include Spanish 21 and pontoon.</td>
<td>Other casino games were inspired by blackjack, including Spanish 21 and pontoon.</td>
<td>5</td>
</tr>
<tr>
<td>HEAVY VERB</td>
<td>After receiving their initial two cards, players have the option of getting a &quot;hit&quot;.</td>
<td>After receiving their initial two cards, players may pursue the option of getting a &quot;hit&quot;.</td>
<td>2</td>
</tr>
</tbody>
</table>

4.3 Results: Transformation strategies

Finally, we look at contrast set classification accuracy across different transformation strategies. In the STATIVE to DYNAMIC direction, the strategy most often classified correctly (56%) is NEW PARTICIPANT. In the reverse direction, the strategy of replacing the DYNAMIC main verb with a COPULA has the highest accuracy (52%). Notably, even the highly stative nature of the copula doesn’t always result in the model recognizing the transformed clause as stative. (Detailed results in Appendix A.)

Both THOUGHT VERB and DESCRIPTIVE VERB for converting DYNAMIC clauses to STATIVE perform poorly. This indicates that verbs of feeling, thinking, and wanting (THOUGHT VERB) and certain verbs like signify, constitute, and include, are not uniformly treated as STATIVE by the classifier. Other descriptive verbs, like contain and resemble do get accurately classified. Some verbs like appear sometimes get classified correctly as STATIVE, as in “All other cards appear as the numeric value”, other times as DYNAMIC, as in “Cards appear either from one or two handheld decks, from a dealer’s shoe, or from a shuffling machine.”

The frequency of transformation strategy in the contrast sets does not correspond to high classification accuracy. In future, we will look at how frequent such constructions are in the training data, and whether their association with stativity labels matches linguistic expectations. We suspect that the different types of stativity may also play a role. For example, the stativity commonly associated with descriptions of mental processes is different from the attributional or predicational stativity often seen with copular constructions.

5 Related work on lexical aspect and verb stativity

Aspects structure is complex and well-studied in the linguistics literature (Vendler, 1967; Comrie, 1976; Moens and Steedman, 1988; Smith, 1991, among many others). Classically, aspectual analysis involves the semantic properties of stativity, telicity, durativity, and iterativity. Croft et al. (2016) expand on this set of properties in their discussion of aspectual annotations within the Rich Event Description framework. Donatelli et al. (2018) propose methods for expanding the Abstract Meaning
### Representation framework with aspectual features

and such aspectual information is a key feature of the Uniform Meaning Representation framework (Van Gysel et al., 2021). In addition, aspectual properties are relevant at both the clause level and the level of individual verbs.

Our current focus is the verb-level property of stativity, sometimes referred to as inherent lexical aspect. Stativity reflects the degree to which a verb represents a static situation versus a situation that reflects some degree of dynamicity. Dynamic verbs typically involve some change of state.

### Computational approaches to stativity

Early approaches to computational analysis of verb stativity employ rule-based approaches based on known linguistic tests for stativity, such as the progressive test. Klavans and Chodorow (1992) produce a type-level stativity rating for English verbs, based on the frequency with which verbs occur in various tenses in the Brown and Reader’s Digest corpora. Dorr and Olsen (1997) treat stativity as one of several aspectual properties derivable from logical representations of verb meaning in the Lexical Conceptual Structure (LCS) framework (Jackendoff, 1983, 1990). Siegel and McKeown (Siegel, 1999; Siegel and McKeown, 2000) use a wide range of linguistic indicators to derive type-level stativity values for English verbs. Friedrich and Palmer (2014a) extend Siegel and McKeown’s work to incorporate distributional features and perform classification in context. Kober et al. (2020) use distributional semantics to classify both stativity and telicity across genres. Falk and Martin (2016) take a more fine-grained approach to lexical aspect classification for French verbs in context, categorizing verbs across a set of 13 different verbal readings. Hermes et al. (2018) take a distributional approach to classifying German verbs for Aktionsart, and (Egg et al., 2019) provide a new annotated corpus and classification experiments for multiple components of aspect for German verbs.

Another important line of research (Govindarajan et al., 2019; Gantt et al., 2022) takes a broader view of event meaning, treating stativity as one of a number of aspectual features which together compose the meaning of an event. Similarly, work on clause-level semantic aspect classification (aka situation entity classification) (Friedrich et al., 2016; Becker et al., 2017; Dai and Huang, 2018) considers stativity as a key semantic property for determining clause-level aspect. Finally, Chen et al. (2021) use a sequence of rules to assign tense and aspect values to both verbal events and event nominals, making use of co-occurrence cues of part-of-speech tags, special lexical items, and semantic configurations that help the classifier select the right shade of aspect for a given situation.

### 6 Conclusions

We apply the contrast set methodology to the task of classifying English verbs in context as stative or dynamic. We see a serious performance degradation on the transformed examples, suggesting the model has not learned a clean decision boundary for stativity. This first analysis suggests a need to more clearly define features that may bias clauses toward stative or dynamic readings.

The study would benefit from more data, across a wider range of text types. We would also like to investigate the effectiveness of contrastive evaluation for other semantic properties, using recently-developed methods for partially-automatic contrast set creation (Li et al., 2020; Bitton et al., 2021; Ross et al., 2021, among others).
7 Acknowledgements

This project is supported, in part, by a grant from the Center for Humanities & the Arts at the University of Colorado Boulder.

References


Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khshahi, Kevin Lin, Jiagming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer


A Classification results by transformation strategy

Table 5 shows the distribution of correct labels assigned to transformed clauses using DYNAMIC to STATIVE strategies. Table 6 shows the same for STATIVE to DYNAMIC transformations.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Test Trans: Size</th>
<th>Test Trans: Correct</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>THOUGHT VERB</td>
<td>55</td>
<td>10</td>
<td>18.18%</td>
</tr>
<tr>
<td>COPULA</td>
<td>48</td>
<td>25</td>
<td>52.08%</td>
</tr>
<tr>
<td>DESCRIPTIVE VERB</td>
<td>47</td>
<td>17</td>
<td>36.17%</td>
</tr>
<tr>
<td>LIGHT VERB</td>
<td>14</td>
<td>4</td>
<td>28.57%</td>
</tr>
<tr>
<td>SEMI-MODAL</td>
<td>5</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>DOWNGRADE TO PPL</td>
<td>2</td>
<td>1</td>
<td>50%</td>
</tr>
<tr>
<td>ORDER</td>
<td>1</td>
<td>0</td>
<td>0%</td>
</tr>
<tr>
<td>Totals</td>
<td>172</td>
<td>57</td>
<td>33.14%</td>
</tr>
</tbody>
</table>

Table 5: Successful DYNAMIC > STATIVE transformation strategies, evaluated by accuracy of correctly identifying the contrast label.

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Test Trans: Size</th>
<th>Test Trans: Correct</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>NEW PARTICIPANT</td>
<td>81</td>
<td>45</td>
<td>55.55%</td>
</tr>
<tr>
<td>INSERT VERB</td>
<td>19</td>
<td>9</td>
<td>47.37%</td>
</tr>
<tr>
<td>BECOMING</td>
<td>13</td>
<td>9</td>
<td>69.23%</td>
</tr>
<tr>
<td>UPGRADE PPL</td>
<td>5</td>
<td>2</td>
<td>40%</td>
</tr>
<tr>
<td>HEAVY VERB</td>
<td>2</td>
<td>1</td>
<td>50%</td>
</tr>
<tr>
<td>Totals</td>
<td>120</td>
<td>66</td>
<td>55%</td>
</tr>
</tbody>
</table>

Table 6: Successful STATIVE > DYNAMIC transformation strategies, evaluated by accuracy of correctly identifying the contrast label.