A *Formidable* Ability: Detecting Adjectival Extremeness with DSMs

Farhan Samir[1], Barend Beekhuizen[2], Suzanne Stevenson[1]
[1]Department of Computer Science
[2]Department of Language Studies and Department of Linguistics
University of Toronto
{fsamir,suzanne}@cs.toronto.edu
barendbeekhuizen@utoronto.ca

Abstract

While distributional semantic models (DSMs) can successfully capture the similarity structure *within* a semantic domain, less is known about their ability to represent abstract semantic properties that hold *across* domains. Such properties can form the basis for abstract semantic classes that are a crucial aspect of human semantic knowledge. For example, the abstract class of extreme adjectives (such as *brilliant* and *freezing*) spans a wide range of domains (here, INTELLIGENCE and TEMPERATURE). Using a model that compares query items to an aggregate DSM representation of a set of extreme adjectives, we show that novel adjectives can be classified accurately, supporting the insight that a cross-domain property like extremeness can be captured in a word's DSM representation. We then use the extremeness classifier to model the emergence of intensifier meaning in adverbs, demonstrating, in a separate task, the effectiveness of detecting this abstract semantic property.

1 Distributional Models and Abstract Semantic Classes

Distributional semantic models (DSMs) are widely used as representations of word-level semantics. However, open questions remain as to precisely which aspects of human semantic knowledge DSMs effectively capture (e.g., Baroni et al., 2014; Hollis and Westbury, 2016; Schnabel et al., 2015; Utsumi, 2020). For example, popular DSMs such as word2vec and GloVe have been shown to predict human ratings of semantic features of objects (Rubinstein et al., 2015; Grand et al., 2018). However, performance is variable across features and object categories (Grand et al., 2018), and in particular, is better for taxonomic properties ('is an animal', 'is a weapon') than for general attributive properties ('is yellow', 'is dangerous') (Rubinstein et al., 2015).

While people may or may not have semantic categories such as "all yellow things", abstract semantic classes are an important part of human linguistic knowledge that should be captured in a computational system. Note that by *abstract* we mean the schematic properties of word meaning, rather than the content-related classes;[1] such properties abstract over commonalities of meaning that may *cross* traditional semantic domains. Consider, e.g., a semantic verb class such as change-of-state (Levin, 1993; Kipper et al., 2008), with members such as *melt* (the TEMPERATURE domain) and *quicken* (SPEED), or relational adjectives (Boleda et al., 2012), including, e.g., *Chinese* (NATIONALITY) or *pulmonary* (BODY-PART).

Much work shows the ability of DSMs to match human knowledge of semantic properties *within* a domain (e.g., Baroni et al., 2014; Pereira et al., 2016; An et al., 2018; Grand et al., 2018), but there is little work, to our knowledge, on whether the similarity structure of a DSM is sensitive to commonalities of abstract properties that hold *across* a variety of semantic domains.[2] Research on vector-based representations of analogy suggests that DSMs may be limited in their ability to represent cross-domain word relations: Rogers et al. (2017) show that cross-domain analogical relations like hypernymy (e.g., *turtle:reptile::salmon:fish*) are significantly harder to solve than within-domain ones (e.g., *Paris:France::Ottawa:Canada*). Lu et al. (2019) make significant progress towards representing such relations, showing that a DSM can form the basis for detecting the cross-domain word

[1]The same distinction between *schematic* and *content* is applied in Paradis (2001); Cruse and Togia (1996). It is also worth noting explicitly that our use of the term 'abstract' in this sense is not to be interpreted as 'not concrete'.

[2]DSMs may, e.g., encode concreteness and valence/arousal/dominance (e.g., Hollis and Westbury, 2016; Hollis et al., 2017), but the former can be viewed as a taxonomic property, and the latter as within the EMOTION domain.

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4112–4125
August 1–6, 2021. ©2021 Association for Computational Linguistics

relations, such as antonymy (*love–hate* in the EMOTION domain, *rich–poor* in FINANCE); however, to achieve this, their method requires additional learning of a relation-specific warping of the distributional semantic space. Our goal here is to see whether the similarity structure of the (original) DSM itself directly captures such cross-domain knowledge.

We focus as a test case on the abstract class of *extreme adjectives*: scalar adjectives that express an extreme value of their scale, such as *brilliant* and *freezing*. Previous computational work on scalar adjectives has focused on assessing their relative ranking within a domain (e.g., learning that *smart* $<$ *brilliant* on the INTELLIGENCE scale) (e.g., Ruppenhofer et al., 2014; Cocos et al., 2018). However, as work in linguistics shows (Cruse, 1986; Paradis, 2001; Morzycki, 2012), extreme adjectives do not simply behave as if they are further along their scale, but rather (as a class) have distinguishing semantic properties. Our goal is to see whether DSMs can capture this cross-domain property of "extremeness". In our first experiment, we demonstrate that we can successfully identify extreme adjectives, across a wide range of domains, on the basis of the information contained in their DSM representations alone.

Our next goal was to show that this ability to detect the abstract property of extremeness would be useful in further tasks. We begin with the novel hypothesis that an adjective's extremeness is a strong predictor of its future use in an intensifier adverb – i.e., in phrases like *staggeringly easy* and *monumentally wrong*. Our second experiment then shows that our classification of extreme adjectives can improve over an existing computational approach (Luo et al., 2019) in a historical prediction task of emerging intensifier meanings.

2 Our Case Study: Extreme Adjectives

Extreme adjectives are a subclass of scalar adjectives that includes words such as *awesome*, *brilliant*, and *freezing*. Like all scalar adjectives, their semantics includes the specification of a scale, such as (for these) DESIRABILITY, INTELLIGENCE, and TEMPERATURE, respectively, along with some position or range on that scale. The distinguishing aspect of extreme adjectives is that they represent an extreme value at one end or the other of the scale (e.g., Cruse, 1986; Morzycki, 2012; Paradis, 2001), a value that is so high/low as to possibly even be considered "off the scale" (Morzycki, 2012). This makes extreme adjectives a good case study for us: their abstract property of "extremeness" holds across a wide variety of semantic scales, and thus crosses individual semantic domains, such as INTELLIGENCE or TEMPERATURE.[3]

Category (example)	Intensification (*very* X)	Endpoint-oriented (*almost* X)	Extreme degree (*absolutely* X)
Non-gradable (*civic*)	?	?	?
Open-scale (*big*)	✓	?	?
Closed-scale (*full*)	✓	✓	✓
Extreme (*huge*)	?	?	✓

Table 1: Examples showing differences in types of modifiers usable across categories of adjectives.

Our goal is to assess whether a DSM can directly capture the similarity among members of this kind of abstract class. We address this question in two ways: with a direct evaluation, testing whether we can classify adjectives as extreme or not, and with an indirect evaluation, which uses our extremeness classifier in a separate task, to show its value in NLP applications.

Our first experiment uses similarity within a DSM as the basis for classification of extreme adjectives. Linguists have long noted that distributional tests can distinguish extreme adjectives from other classes (e.g. Cruse, 1986), as in the following examples (elaborated in Table 1).[4]

1. Martha is ?very/?almost/absolutely **ecstatic**. [Extreme]

2. Martha is very/?almost/?absolutely **grateful**. [Not Extreme]

Here, we propose an approach that replaces similarity along manual distributional tests with similarity within a general DSM. Note that such an approach is not a priori guaranteed to succeed. First, the manually-identified probes are specifically chosen to highlight the distinguishing properties; looking at representations derived from all of a word's contexts may mean that the useful signal about what makes an adjective extreme is drowned in "noise"

[3]The difference between a semantic domain and a semantic scale is not crucial here; the important point is that "extremeness" is a property that crosses what are typically thought of as more narrowly-defined semantic areas.

[4]Note that the versions with "?" in (1), (2), and Table 1 are less felicitous and require additional facilitating context, rather than being outright ungrammatical; this is not unusual for distributional tests of semantic classes.

(with respect to the property of extremeness). Second, we're looking for an abstract property that crosses semantic domains – i.e., we're not asking if EMOTION adjectives are more similar to each other than to INTELLIGENCE adjectives; we're asking whether extremes in both domains (*ecstatic* and *brilliant*) are more similar to each other than to non-extreme adjectives.

In our second experiment, we perform a downstream task that uses our classification of an adjective as extreme or not. Here we study the historical emergence of intensifier meanings of adverbs (e.g., *monumentally* coming to mean 'very', as in *monumentally wrong*), for which in-domain similarities in DSMs have previously been found to model the phenomenon with some accuracy (Luo et al., 2019). Using our cross-domain model of extremeness, we operationalize our novel linguistic insight that adjectival extremeness (*monumental*) is inextricably linked to the emergence of intensifier senses (*monumentally wrong*).

To preview our results, we find that our method of using a DSM to classify adjectives as extreme or not substantially improves over a statistical method drawing on linguistically-devised distributional tests. That is, we see that the general similarity space of a DSM can be used as an effective replacement for similarity with respect to manually-identified probes in the detection of an abstract semantic class. Moreover, in our experiments on intensifier emergence, we find that our identification of the cross-domain property of extremeness also shows improvement over a more standard application of DSMs that assesses in-domain similarities. This pair of experiments thus provides evidence that a DSM can successfully capture an abstract class defined by cross-domain similarity.[5]

3 Classifying Extreme Adjectives

Here we propose an approach for identifying extreme adjectives using a DSM. Other work in computational linguistics has considered automatic means for placing scalar adjectives in the appropriate relative position along their scale (e.g. Sheinman et al., 2013), including approaches using DSMs to do so (e.g. Kim and de Marneffe, 2013; Sharma et al., 2017). Such methods do not distinguish extreme adjectives as a special set across

[5]All code and data are available at https://github.com/smfsamir/detect-adjectival-extremeness.

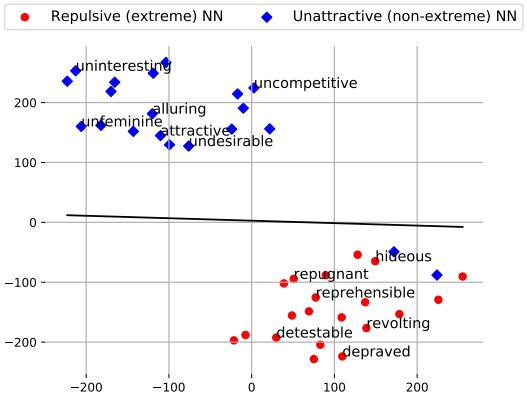

Figure 1: Nearest neighbours of *repulsive* (extreme) and *unattractive* (non-extreme), visualized with t-SNE (Maaten and Hinton, 2008) on word2vec embeddings (Mikolov et al., 2013).

semantic domains, but rather assess them as expressing a higher/lower level compared to other adjectives within the domain – e.g., *brilliant* is a higher degree of INTELLIGENCE than *smart*. Our approach instead focuses on extremeness as a categorical property of adjectives that is independent of any particular scale (following Cruse, 1986; Paradis, 2001; Morzycki, 2012, among others).

As set out earlier, our hypothesis is that the similarity space of a DSM can capture the (cross-domain) similarity of the members of an abstract semantic class. Fig. 1 illustrates this intuition: the nearest neighbors of the extreme adjective *repulsive* are other extreme adjectives (*detestable*, *revolting*); conversely, the non-extreme counterpart *unattractive* has other non-extreme nearest neighbours (*undesirable*, *uninteresting*). Importantly, the two sets of nearest neighbours show a clear separability of the extreme adjectives and non-extreme adjectives, suggesting that the contextual distribution of an adjective, as represented in its word embedding, contains useful information for classifying whether it is extreme.

Reflecting this intuition, we propose to identify extremeness of adjectives using a prototype approach, in which we classify each adjective by comparing its vector to the average vector (a "prototype") of a set of extreme adjectives within a DSM. While such an approach has been used previously to characterize words *within* a semantic domain (e.g., EMOTION in the case of Xu et al. (2020), and each of a variety of domains in An et al. (2018)), here we test whether such a prototype vector can abstract over the individual semantic domains to

capture the cross-domain property of extremeness. Classification will be successful to the extent that vectors of adjectives that share the abstract property of extremeness are sufficiently similar to serve as an informative prototype.

3.1 Dataset

We collect a dataset of extreme and non-extreme adjectives in English for training and evaluating a supervised classifier. We started with a set of 54 adjectives identified as extreme (Morzycki, 2012; Paradis, 2001; Huttenlocher et al., 1971; Cruse, 1986; Lassiter, 2017). We then added extreme adjectives from human-annotated datasets of adjectival intensity (Cocos et al., 2018; Wilkinson and Tim, 2016; de Melo and Bansal, 2013; Ruppenhofer et al., 2014). For each adjective, these datasets specify its scale and its human-rated range of intensity values. For each of the scales of the 54 previously-identified extreme adjectives, we gathered all further adjectives tied with or ranked above extreme adjectives in their intensity value ($N = 17$). After filtering out 3 adjectives with frequency less than 0.5 per million, we obtained a total of $N = 68$ extreme adjectives that cover a diverse set of adjectival scales, such as DESIRABILITY (*sensational*), INTELLIGENCE (*moronic*), and SIZE (*gigantic*). We then match each extreme adjective in our dataset with a non-extreme adjective matched for frequency in COCA (Corpus of Contemporary American English; 1B words; Davies, 2009). (Since word embeddings have been shown to encode frequency (e.g., Mu and Viswanath, 2018), it is important to control for this.) To avoid including extreme adjectives, we exclude any words that appear as the most intense adjective in the above datasets. Our non-extreme set includes 68 adjectives, both scalar and non-scalar; see data in the GitHub repository.

3.2 Methods to predict extremeness

Our goal is to see whether DSMs can support the classification of an abstract class that has been previously identified through manually-selected distributional tests. Thus we compare two approaches to identifying extreme adjectives: one implements a corpus-based measure to capture manually-identified distributional tests developed by semanticists (cf. Table 1), while the other uses a DSM to capture the cross-domain similarity of extreme adjectives. For both, we propose a method for deriving the probability $p(c|a)$ of the class $c \in \{\text{Extreme}, \text{Non-extreme}\}$ for an adjective a.

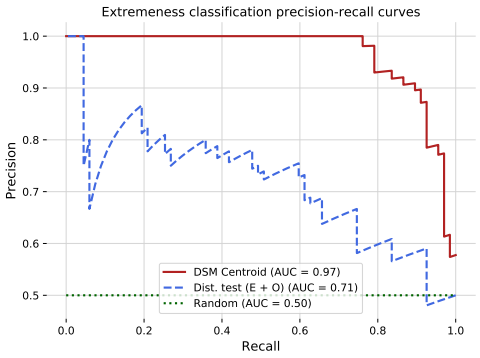

Figure 2: Results for classifying extreme adjectives.

3.2.1 The DISTRIBUTIONAL TEST method

Here we define a method based on the "probe" words specified in distributional tests of extremeness. Specifically, our approach looks at patterns of extreme adjectives that are readily assessed in a corpus: (1) frequent modification by extreme degree modifiers (Cruse, 1986); and (2) resistance to modification by endpoint-oriented modifiers and *very* (Morzycki, 2012).[6] The extreme degree modifiers we consider are E = {*absolutely*, *totally*, *simply*, *positively*, *downright*, *outright* }, while the other modifiers are O = { *almost*, *slightly*, *very* }. Formally, we measure the association with each of these sets of probes using normalized pointwise mutual information (NPMI).[7] For example, we assess the NPMI_E between an adjective a and (all) extreme degree modifiers e that immediately precede it in a sentence, as:

$$\text{NPMI}_E(e, a) = \log\left(\frac{p(e,a)}{p(e)p(a)}\right) \cdot \frac{1}{\log p(e,a)}$$

where e stands for any extreme degree modifier in E (i.e., we derive the probabilities above from pooled counts in COCA over a co-occurring with any $e \in E$). We compute the association NPMI_O with all $o \in O$ analogously. Using NPMI_E and NPMI_O as features, we train a logistic regression classifier with L2 regularization to obtain an estimation of $p(c|a)$, where $c \in \{\text{Extreme}, \text{Non-extreme}\}$. Combined with a probability threshold, the estimation of $p(c|a)$ allows us to predict whether an adjective is extreme.

[6] Other tests (such as prosodic cues or hyperbole, Morzycki, 2012), are difficult to identify in a written corpus.

[7] Normalizing mitigates the frequency bias known to impact PMI (Bouma, 2009; Jurafsky and Martin, 2014).

Classifier	Prec.	Recall	F_1
DISTRIB. TEST	0.76	0.66	0.71
DSM CENTROID	**0.90**	**0.91**	**0.90**

Table 2: Classification using a 0.5 threshold.

3.2.2 The DSM CENTROID method

To use a DSM to classify an adjective as extreme, we need to abstract away from the information about the particular scale or semantic domain that is captured in a word vector. That is: we want the model to recognize that *exquisite* is more like *huge*, *destitute*, and *repulsive* than like *big*, *poor*, and *unattractive*. As noted above, we draw on prototype approaches by classifying a novel adjective on the basis of how similar it is to the aggregate representation of a set of extreme adjectives. Analogously to the approach above, we use this similarity as a single feature in a logistic regression classifier to obtain an estimate of the probability of extremeness $p(c|a)$. For our experiments, we use the 300-dimensional pretrained word2vec embeddings of Mikolov et al. (2013).

3.3 Results

Because of the small size of our dataset ($N = 136$), we use a leave-one-out cross-validation procedure to evaluate the two methods. We evaluate performance in two ways: using precision-recall curves (shown in Fig. 2) and as Precision, Recall, and F_1 scores for the Extreme class using a classification threshold of 0.5 (shown in Table 2).

Both DSM CENTROID (AUC = .97) and DISTRIBUTIONAL TEST (AUC = .78) perform better than chance (AUC = .50). The DSM CENTROID method furthermore substantially outperforms the DISTRIBUTIONAL TEST method (AUC = .97 vs. AUC = .78; F_1 = .90 vs. F_1 = .71). This result provides strong evidence for our hypothesis that the similarity structure of a DSM can effectively capture an abstract, cross-domain property such as extremeness; in fact, at least in this case, it can do so better than a statistical corpus-based model based on manually-identified linguistic tests.

One concern with the leave-one-out approach is that there may be items in the training set (e.g., *massive*) that share semantic properties other than extremeness with the held-out item (e.g., *huge* – in this case, both are SIZE adjectives). To control for this confound, we conduct a variant of a k-fold cross-validation procedure, in which the folds consist of clusters of semantically similar adjectives, with the aim that the training data does not include adjectives from the same domain as the test items. Even in this controlled analysis, we find that the DSM CENTROID method robustly outperforms the DISTRIBUTIONAL TEST method. See Appendix A for details.

Classifier	Errors
DSM CENTROID	destitute, freezing, terrified; *runaway, memorable*
DISTRIB. TEST	gigantic, colossal, mammoth, gargantuan, immense; *identical, inconvenient, conventional, opposite*
Both	major, obese, microscopic; *creepy*

Table 3: Sample of errors made by each method, and both. False negatives; *false positives*.

3.4 Discussion

Given that the DSM CENTROID method uses information from all of a word's contexts (as captured by the DSM), in contrast to the DISTRIBUTIONAL TEST approach that uses hand-picked tests to reveal behavior relevant to extremeness, it is worth exploring why the former method performs so much better. Here we study the errors made by the two methods when using a 0.5 probability threshold for the classifiers. We focus on the sets of errors that are exclusive to each model, as shown in Table 3; a complete list of errors is available in the data provided in the GitHub repository.

One possibility is that distributional tests will fail to identify extreme adjectives when they only infrequently co-occur with the distributional probes; the fact that an extreme adjective *can* be modified by, e.g., *absolutely* does not entail that these will actually co-occur in the corpus. The false negatives for the DISTRIBUTIONAL TEST method reflect this: it misses out on extreme adjectives that do not frequently co-occur with extreme degree modifiers. We see, for instance, that many extreme adjectives in the domain of SIZE are misclassified due to a low NPMI_E score – a significant error since extreme adjectives of SIZE are often presented as typical members of the class (Cruse, 1986; Morzycki, 2012). The DISTRIBUTIONAL TEST method also produces a considerable number of false positives. This reflects another shortcoming of distributional tests, as it has been noted that extreme degree modifiers

can be used with adjectives that are *not* extreme (cf. *absolutely inconvenient*, *totally identical*; Paradis, 2001; Morzycki, 2012).

The cases where DSM CENTROID fails – false negatives and false positives – are mostly of a different nature. Among its errors are many words that have *both* extreme and non-extreme senses (*freezing*, *destitute*, *microscopic*, *obese*, *runaway*, and *major*). For example, *destitute* can mean 'devoid of', and *freezing* can refer to the actual transitioning from liquid to solid, resulting in a non-extreme classification. Conversely, *runaway* is misidentified as extreme due to extreme uses such as *runaway success*. Such polysemies highlight an issue for any method that draws on data from a corpus that is not sense-annotated – recognizing an adjective as extreme depends on the word being predominantly used in the expected sense.

Despite this challenge, our results align with the analysis of extreme adjectives as a distinct adjectival class (Cruse, 1986; Paradis, 2001; Morzycki, 2012). Members of this category can be automatically identified using similarity to an aggregate vector of (known) extreme adjectives, even when those exemplars cross a wide range of semantic domains. Importantly, while Morzycki (2012) argued that extreme adjectives are always felicitous with extreme degree modifiers such as *absolutely*, this does not entail that the association will necessarily manifest, even in a large corpus. Instead, we leverage the broad distributional tendencies captured by DSMs to determine semantic class membership probabilistically.

Our good performance in this task required only a single centroid representation of our semantic class of interest – importantly, it did not require an opposing negative centroid (as in, e.g., An et al., 2018) nor a relation-specific warping of semantic space as in Lu et al. (2019). In future work, we hope to use recent developments in applying influence functions for NLP (e.g., Brunet et al., 2019) to discover the contexts that enable DSMs to represent such abstract properties effectively.

4 Predicting novel intensifiers

Having shown that extremeness can be classified on the basis of similarity within a DSM, we now show that this method is sufficiently informative to guide a separate task – that of modeling the emergence of novel intensifiers. Intensifiers are adverbs such as *staggeringly* in *staggeringly easy*, or *monumentally* in *monumentally wrong*, that give force to the modified adjective without conveying their manner meaning – i.e., saying some task is *staggeringly easy* means that it is extremely easy, not that it is easy in a staggering way. Manner adverbs can gain such an intensifying sense through a continual, and frequent, process of semantic change (Bolinger, 1972; Bennett and Goodman, 2018; Luo et al., 2019). However, it is an open question why certain manner adverbs are more likely to do so than others.

Our hypothesis is that it is the abstract property of extremeness that facilitates this meaning shift – intuitively, *staggeringly easy* and *monumentally wrong* can readily be paraphrased as *off-the-scale easy* and *off-the-scale wrong* because the underlying adjectives (*staggering* and *monumental*) denote extreme values. Thus, an ability to detect extremeness should support an approach to prediction of emergence of intensifier senses: i.e., the adverbs derived from extreme adjectives should be those that are likely to become intensifiers.

We contrast our hypothesis with that of Luo et al. (2019), who propose that intensifier usage can arise with any adverb through a more general process of semantic bleaching (Traugott and Dasher, 2001). Specifically, Luo et al. (2019) argue that when adverbs modify semantically similar adjectives (*conspicuously evident*), the redundancy of the modifying adverb leads to its interpretation as intensifying the content of the adjective, and the actual manner component of the adverb is bleached over time. Luo et al. (2019) capture this insight with a measure (described below) of the within-domain similarity of adverbs to the adjectives that they are found to modify frequently.[8]

We follow Luo et al. (2019) in adopting a semantic bleaching account, but instead propose that it is the abstract property of extremeness that is the focus of the bleaching process: On our account, intensification arises when the particular scale (EMOTION or INTELLIGENCE) of an extreme adverb (an adverb derived from an extreme adjective) is backgrounded, and eventually lost, while the abstract property of extremeness remains the key part of the meaning.

Our approach is compatible with that of Luo et al. (2019), but moreover *explains* why certain adverbs are more likely to modify similar adjec-

[8] By "within-domain" we mean a standard topical/domain-level similarity of meaning, such as *conspicuously* and *evident* conveying 'ease of observation'.

Semantic category	Entries
The world > Relative properties > Quantity > Greatness of quantity/ amount/ degree > hugely	**colossally**, **monumentally**, hugeously, thumpingly, pyramidically
In respect of quantity > Greatly/very much > extremely/exceedingly > remarkably/ extraordinarily	**markedly**, **exceptionally**, **noticeably**, **pronouncedly**, **prominently**
Consciously, knowingly > In accordance with truth, truly > in fact, actually	**literally**, **absolutely**, **objectively**, essentially, factually

Table 4: HTE categories with a sample of entries. Bolded words are attested after 1800.

tives in the first place: Adverbs that are derived from an extreme adjective are likely to modify similar adjectives precisely because they are *not* redundant – they contribute the salient meaning of 'extremeness' over and above the expression of the scale already expressed in the modified adjective. Thus we suggest that it is primarily the abstract, cross-domain property of scalar extremeness that drives the emergence of intensifying meanings of adverbs, rather than their patterns of co-occurrence with adjectives that are semantically similar. In this section, we support our claim with a modelling experiment that contrasts the within-domain similarity measure of Luo et al. (2019), with our cross-domain similarity measure for identifying extremeness.

4.1 Methods and materials

To compare these hypotheses, we perform a historical prediction task: predicting the emergence of an intensifier sense of an adverb based on data in the decades prior to the sense's attestation date. Note that this differs from Luo et al. (2019), who performed statistical analyses over time, including both before and after emergence of the intensifying senses. Consequently, we cannot adopt the full dataset of Luo et al. (2019) since it includes adverbs whose intensifier sense emerged prior to 1800 – where historical corpus data is scarce.

Intensifiers. We start with a set of 69 intensifiers R that have their first date of attestation of an intensifier sense after 1800 according to the Historical Thesaurus of English (HTE; Kay et al., 2017). (These include 45 from Luo et al. (2019) that meet this criterion.) These adverbs come from a wide variety of semantic categories, as illustrated in Table 4. We restrict the attestation date to post-1800 so that we can draw on a large historical corpus with sufficient data and robust word embeddings.

Prediction timeframe. Because we aim to predict the *emergence* of novel intensifiers, rather than classify existing ones, the model should not use data that contains the target adverb in its intensifier meaning. In order to use the same time span for all adverbs and include a reasonable quantity of data, we take data for each adverb from the 3 decades prior to its use as an intensifier. That is, for each adverb r attested within decade $d_A(r)$, we use corpus data from the 3 decades $T_r = \{d_A(r) - 3, d_A(r) - 2, d_A(r) - 1\}$.

Matched control adverbs. Luo et al. (2019) presented a set of 178 control adverbs C that did not develop an intensifying meaning. We match each $r \in R$ with the $c \in C$ that has the most similar frequency to r in the time period T_r, and use data for c from the same timeframe T_r.

Historical corpus data. With smaller corpora being too sparse, we use historical data (years 1800–1999) from the Google N-grams corpus (English, version 2, Michel et al., 2011), drawing on the diachronic skip-gram word embeddings of Hamilton et al. (2016), and syntactic annotations from Goldberg and Orwant (2013). Note that the embeddings are formed per-decade, to allow for sufficient training data. In our primary analyses, we use the embeddings of the adjectival bases of the adverbs, due to sparsity of the adverb embeddings themselves. Due to missing embeddings for some of the items, we had to remove 28 items, leaving us with 52 intensifiers and 58 control adverbs.

4.2 Features predicting intensifier emergence

Here we describe the two features we are investigating as predictive of the development of adverbs into intensifiers, our EXTREMENESS feature and the SIMADJMOD feature of Luo et al. (2019).

EXTREMENESS. To test our hypothesis that extreme adverbs are likely to gain an intensifying sense, we adapt the DSM CENTROID measure from Section 3. For each adverb q to be classified, we first find its adjectival basis a_q (e.g., *insane* for *insanely*), and then obtain the embedding $\mathbf{a_q}$ as the average of the diachronic embeddings $\mathbf{a_q}(\mathbf{t})$ for each decade $t \in T_q$. We then obtain the likelihood that a_q is in class c – where $c \in \{\text{Extreme}, \text{Non-extreme}\}$ – based on its proximity to the extremeness centroid. We obtain the extremeness centroid using the extreme adjectives in Section 3, averaging their diachronic embeddings

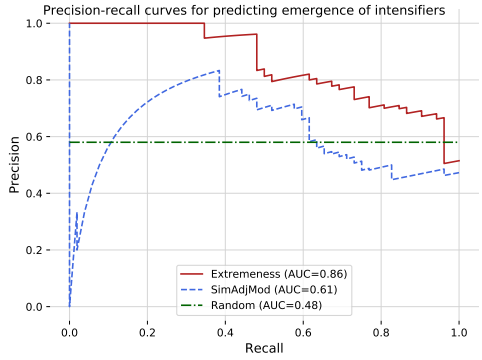

Figure 3: Results for historical prediction task.

over the time period T_q. While this is a dataset of *currently* extreme adjectives, we believe it can be reliably used for the time period covered, as extremeness seems to be a relatively stable property.

SIMADJMOD. Luo et al. (2019) formalize their hypothesis with a measure, SIMADJMOD, of the similarity between an adverb and the adjectives it modifies. For a query adverb q, the measure is computed as the average semantic similarity between q (e.g., *conspicuously*) and the set of all adjectives modified by q (e.g., *evident*). We adapt this measure with an approach analogous to the EXTREMENESS measure: we find the word embeddings of the adjective a_q that is the basis of q and of all adjectives modified by q within T_q. Then, SIMADJMOD(q) is computed as the odds-weighted average of the cosine similarity of $\mathbf{a_q}$ with each modified adjective's embedding (see Luo et al. (2019) for details on the weighting function).

It is worth noting that using adjective embeddings contrasts with Luo et al. (2019), who use the adverb embeddings themselves. We adopted our approach due to the absence of many adverbs in the diachronic embeddings. In their statistical analyses (in contrast to our prediction task), Luo et al. (2019) can use embeddings from a 150-year time span, reducing the sparsity of adverb embeddings. We perform an additional experiment in which we follow their set-up more closely, reported in Appendix B. As in our main results below, the EXTREMENESS feature comes out as a strong explanatory factor, although its improvement over the SIMADJMOD feature is much greater in the core historical prediction task. We discuss the implications in Appendix B.

4.3 Results and Discussion

We fit a separate logistic regression classifier for each feature, EXTREMENESS and SIMADJMOD, and use leave-one-out cross-validation. We report precision-recall curves using the $p(c|q)$ estimates from the regression models (where $c \in \{\text{Intensifier}, \text{Control}\}$), and perform error analysis using a 0.5 threshold on $p(c|q)$. The results are presented in Figure 3. We observe that both measures perform well above chance, with the EXTREMENESS feature (AUC $= 0.86$) outperforming the SIMADJMOD feature (AUC $= 0.61$).[9] To understand the difference in performance between the two features, we turn to an analysis of the errors.

One situation that should distinguish the use of EXTREMENESS from SIMADJMOD is in predicting the emergence of intensifier meanings for adverbs that are infrequent: an uncommon extreme adverb will still be extreme, but its low frequency will prevent it from frequent co-occurrence with similar adjectives. We find evidence for this in the errors: EXTREMENESS, but not SIMADJMOD, correctly predicts the emergence of an intensifier meaning for the infrequent adverbs *monumentally*, *colossally*, *frighteningly*, *staggeringly*, and *thunderingly*. This supports our hypothesis that it is not the frequent co-occurrence (with similar adjectives) that primarily drives intensifier emergence, but the semantic property of extremeness.

However, while many novel intensifiers are based on adverbs derived from extreme adjectives, not all of them are; adverbs that have an epistemic function (e.g., *actually*, *really*, *honestly*) also form a frequent source of intensifiers (Bolinger, 1972; Traugott and Dasher, 2001). We indeed observe that the EXTREMENESS feature fails to predict the emergence of intensifiers such as *honestly* and *prominently*, cases of adverbs that were not derived from extreme adjectives, while the SIMADJMOD feature captures these cases.

Overall, the results of this experiment support our claim: the cross-domain semantic property of extremeness is highly effective in predicting whether an adverb will gain an intensifying function. While the account of Luo et al. (2019) provides substantial empirical coverage of the pattern of historical development, it does not explain why particular adverbs should modify similar adjectives in the first place. Our account subsumes theirs

[9] Including both features in the classifier does not improve performance over the EXTREMENESS feature alone.

by showing that if a manner adverb frequently co-occurs with a similar adjective, the adverb is likely derived from an extreme adjective (with exceptions, such as the epistemic adverbs); the adverb is then easier to re-interpret as an intensifier by maintaining the feature of extremeness while bleaching its domain-specific semantics.

5 Conclusion

In this paper, we looked at the question of whether distributional semantic models (DSMs) can capture abstract semantic properties of word classes. While it is well known that such models can capture within-domain similarity (e.g., *tall* being similar to *high* and *large*, all pertaining to the scalar attribute of SIZE), here we study whether DSMs also encode abstract similarities that hold *across* domains.

Our case study focuses on extremeness (Cruse, 1986; Morzycki, 2012; Paradis, 2001), the adjectival property of expressing values so high or low that they can be considered "off the scale" (Morzycki, 2012). Extreme adjectives thus share the abstract property of "extremeness" on a scale, independently of the particular scale involved. Our experiments demonstrate that the abstract property of extremeness can be identified using similarity of an adjective to an extreme "prototype" in a DSM, thus supporting the claim that the similarity structure of a DSM can encode such cross-domain classes.

We believe our study contributes to the rapprochement of computational methods and linguistic theory (cf. Boleda, 2020): both the traditional application of distributional tests, as well as the more open-ended application of DSMs, rely on distributional patterns. Since those patterns are more opaque in a DSM, a fruitful next step would be to reverse-engineer the distributional patterns that are indicative of abstract semantic class membership and thereby 'give back' an interpretable set of class membership indicators to semantic theory.

6 Acknowledgments

This research was supported by NSERC grant RGPIN-2019-06917 to Barend Beekhuizen and by NSERC grant RGPIN-2017-06506 to Suzanne Stevenson.

References

Jisun An, Haewoon Kwak, and Yong-Yeol Ahn. 2018. SemAxis: A lightweight framework to characterize domain-specific word semantics beyond sentiment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247.

Erin D. Bennett and Noah D. Goodman. 2018. Extremely costly intensifiers are stronger than quite costly ones. *Cognition*, 178:147–161.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*.

Gemma Boleda, Sabine Schulte im Walde, and Toni Badia. 2012. Modeling regular polysemy: A study on the semantic classification of catalan adjectives. *Computational Linguistics*, 38(3):575–616.

Dwight Bolinger. 1972. *Degree words*. The Hague: Mouton.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2019. Understanding the origins of bias in word embeddings. In *Proceedings of the 36th International Conference on Machine Learning*, pages 803–811. PMLR.

Anne Cocos, Veronica Wharton, Ellie Pavlick, Marianna Apidianaki, and Chris Callison-Burch. 2018. Learning scalar adjective intensity from paraphrases. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1752–1762.

D Alan Cruse and Pagona Togia. 1996. Towards a cognitive model of antonymy. *Lexicology*, 1(1):113–141.

David Alan Cruse. 1986. *Lexical semantics*. Cambridge University Press.

Mark Davies. 2009. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2):159–190.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of English books. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, pages 241–247.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2018. Semantic projection: recovering human knowledge of multiple, distinct object features from word embeddings. *arXiv preprint arXiv:1802.01241*.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501.

Geoff Hollis and Chris Westbury. 2016. The principals of meaning: Extracting semantic dimensions from co-occurrence models of semantics. *Psychonomic Bulletin & Review*, 23(6):1744–1756.

Geoff Hollis, Chris Westbury, and Lianne Lefsrud. 2017. Extrapolating human judgments from skip-gram vector representations of word meaning. *The Quarterly Journal of Experimental Psychology*, 8(70):1603–1619.

Janellen Huttenlocher, E Tory Higgins, and Herbert H Clark. 1971. Adjectives, comparatives, and syllogisms. *Psychological Review*, 78(6):487.

Dan Jurafsky and James H Martin. 2014. Speech and language processing. vol. 3.

C. Kay, J. Roberts, M. Samuels, and I. Wotherspoon. 2017. *The Historical Thesaurus of English, version 4.21*. University of Glasgow, Glasgow, UK.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42(1):21–40.

Daniel Lassiter. 2017. *Graded modality: Qualitative and quantitative perspectives*. Oxford University Press.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Hongjing Lu, Ying Nian Wu, and Keith J Holyoak. 2019. Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences*, 116(10):4176–4181.

Yiwei Luo, Dan Jurafsky, and Beth Levin. 2019. From insanely jealous to insanely delicious: Computational models for the semantic bleaching of english intensifiers. In *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, pages 1–13.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(85):2579–2605.

Gerard de Melo and Mohit Bansal. 2013. Good, great, excellent: Global inference of semantic intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014):176–182.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Neural and Information Processing System (NIPS)*.

Marcin Morzycki. 2012. Adjectival extremeness: Degree modification and contextually restricted scales. *Natural Language and Linguistic Theory*, 30(2):567–609.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Carita Paradis. 2001. Adjectives and boundedness. *Cognitive Linguistics*, 12(1):47–65.

Francisco Pereira, Samuel Gershman, Samuel Ritter, and Matthew Botvinick. 2016. A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, 33(3-4):175–190.

Anna Rogers, Aleksandr Drozd, and Bofang Li. 2017. The (too many) problems of analogical reasoning with word vectors. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (* SEM 2017)*, pages 135–148.

Dana Rubinstein, Effi Levi, Roy Schwartz, and Ari Rappoport. 2015. How well do distributional models capture different types of semantic knowledge? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 726–730.

Josef Ruppenhofer, Michael Wiegand, and Jasper Brandes. 2014. Comparing methods for deriving intensity scores for adjectives. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 117–122.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of*

the 2015 Conference on Empirical Methods in Natural Language Processing, pages 298–307.

Raksha Sharma, Arpan Somani, Lakshya Kumar, and Pushpak Bhattacharyya. 2017. Sentiment intensity ranking among adjectives using sentiment bearing word embeddings. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 547–552.

Vera Sheinman, Christiane Fellbaum, Isaac Julien, Peter Schulam, and Takenobu Tokunaga. 2013. Large, huge or gigantic? identifying and encoding intensity relations among adjectives in wordnet. *Language resources and evaluation*, 47(3):797–816.

Elizabeth Closs Traugott and Richard B Dasher. 2001. *Regularity in semantic change*. Cambridge University Press.

Akira Utsumi. 2020. Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, 44(6).

Bryan Wilkinson and Oates Tim. 2016. A gold standard for scalar adjectives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2669–2675.

Aotao Xu, Jennifer Stellar, and Yang Xu. 2020. Prototype theory and emotion semantic change. In *Proceedings CogSci*.

A Extremeness classification: supplementary analysis

In this supplementary analysis, we report on an experiment in which we control for the potential confound described at the end of Section 3.3; that is: some training set items frequently share domain-specific features with the held out test item (e.g., *massive* and *colossal* are in the training set while *huge* is held out). To validate that the DSM CENTROID method is detecting the abstract property of extremeness, rather than simply tapping into such domain-specific similarities, we set up the evaluation procedure in a way that minimizes the possibility that domain-specific similarities explain the results. We do so by aiming to exclude adjectives from the training set that are in the same domain as a test adjective.

A.1 Methods and materials

Because we do not have ground-truth domain labels for our extreme adjectives (e.g., {*huge*, *colossal*, *mammoth*} $\in$ SIZE), we require another method for grouping the adjectives into domains. As an approximation of domain groupings, we use an unsupervised clustering algorithm on the extreme adjectives, expecting that items belonging to the same cluster will belong to the same domain. With these k domain-based clusters, we can then train on $k-1$ of the clusters while testing on the held-out cluster.

Concretely, we perform a variant of k-fold cross-validation, called leave-one-cluster-out cross-validation, following Utsumi (2020), where the folds are determined through k-means clustering of our dataset of extreme adjectives. For binary classification of extreme adjectives, we train the DSM CENTROID and DISTRIB. TEST methods, described in Section 3.2, on extreme adjectives in $k-1$ of the clusters as well as their frequency-paired non-extreme adjectives; see Section 3.1 for details of this pairing. We then test our binary classifiers on the extreme adjectives in the k-th cluster along with their paired non-extreme adjectives. We repeat this procedure k times so that each cluster is tested once.

A.2 Results

The leave-one-cluster-out cross-validation procedure is dependent on the number of clusters k used for k-means clustering. We perform the procedure for a wide range of $k \in [2, \ldots, 12]$ to ensure that the results are robust to the choice of k. Table 5 (below) displays subsets of the clusters obtained with k-means clustering for $k = 5$, qualitatively demonstrating that the clusters are centered on domain-specific features. The cross-validation results on the full range of k values are shown in Fig. 4.

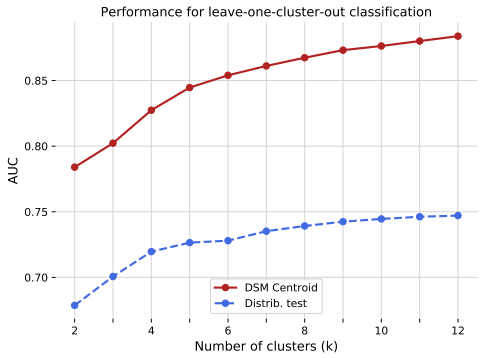

Figure 4: Results for extremeness classification using the leave-one-cluster-out cross-validation procedure. We vary the number of clusters from $k \in [2, \ldots, 12]$.

We can see that the DSM CENTROID performance here (AUC $\in [0.78, \ldots, 0.88]$) is lower than in Section 3.3 (AUC=0.97), indicating that within-domain similarity was partly responsible for the result in that set-up. However, even in this more controlled set-up, the DSM CENTROID method robustly outperforms the DISTRIBUTIONAL TEST method (AUC $\in [0.67, \ldots, 0.74]$, which is similar to the AUC of 0.71 for this method in Section 3.3). We take the results in this controlled analysis as further evidence that the cross-domain property of extremeness can be detected from DSM vectors.

B Intensifier prediction: supplementary analysis

In Section 4, we used a historical prediction task to test two contrasting accounts of how intensifiers emerge in the lexicon. Specifically, we developed classifiers, based on diachronic DSMs, to predict whether an adverb would acquire an intensifying sense based on historical corpus data *prior* to the sense's attestation date. By restricting to corpus data prior to attestation dates, we ensured that the DSM vectors reflected manner adverb usage (*monumentally sized*) rather than intensifier usage (*monumentally wrong*). That is, we minimized the risk that the DSM vectors smuggled in information of whether an adverb was already an intensifier.

This restriction, however, also results in two data

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
ecstatic	brainless	gargantuan	penniless	mesmerizing
phenomenal	amateurish	miniscule	destitute	resplendent
exceptional	moronic	mammoth	miserable	immaculate
terrific	inane	colossal	filthy	breathtaking
outstanding	idiotic	monstrous	obese	brilliant

Table 5: Samples of clusters from k-means clustering on extreme adjectives with $k = 5$.

sparsity problems. First, there are relatively few adverbs that gained an intensifying sense after the 1800s – the time period for which we have historical corpus data – so we were required to use a smaller sample of intensifiers. Second, many adverbs gained intensifying senses towards the earlier parts of the 19th century, for which historical corpora are significantly smaller than in later decades. As described in Section 4, the latter issue required us to compute the SimAdjMod measure differently than Luo et al. (2019): We used embeddings of the adjectival basis (*insane*) rather than the adverb itself (*insanely*), since many adverbs were missing in the historical embeddings.

In this section, we perform the prediction task introduced in Section 4 without restricting to corpus data prior to attested dates of intensifier senses. While lifting this restriction weakens the explanatory account of how intensifiers emerge in the lexicon (due to the aforementioned methodological concern), it alleviates the problem of data sparsity and aligns more closely with the experimental set-up of Luo et al. (2019), thereby making the results more directly comparable. This experiment thus serves as a follow-up test for our hypothesis that EXTREMENESS is a strongly predictive feature of whether intensifying meanings emerge for manner adverbs.

The experiment here is similar to the one reported in Section 4, but there are three significant differences, all introduced to align better with the setting of Luo et al. (2019): First, we use the majority of intensifiers ($N = 185$ out of 250) and control adverbs ($N = 152$ out of 178) published by Luo et al. (2019) – specifically, we use all of their adverbs that have available word embeddings, as opposed to only intensifiers attested within the 1800s to the present. Second, we use a prediction timeframe of $T = \{1850, 1860, \ldots, 1990\}$ for all queries, as opposed to a timeframe of 3 decades parameterized by an intensifier's attested date. Third, we use the adverb embeddings in computing the SIMADJMOD measure, as opposed to the embeddings of the adverbs' adjectival bases used in Section 4.

B.1 Methods and materials

Corpus. The binary classification – as having an intensifier sense or not – is conducted with corpus data spanning 15 decades from the 1850s to the 1990s from the Google N-grams and syntactic N-grams corpora. We used diachronic skip-gram word embeddings as in Section 4.

Evaluation dataset. Our evaluation dataset consists of the 185 intensifiers and 152 control-adverbs from the Luo et al. (2019) dataset that were left after discarding, similarly to their modelling set-up, 65 intensifiers and 26 control-adverbs with missing word embeddings.

B.2 Features for classification of intensifiers

For each query adverb q, we compute the two features (EXTREMENESS and SIMADJMOD) as follows:

EXTREMENESS. We calculate the cosine similarity of the adjectival basis embedding $\mathbf{a_q(t)}$ to the extremeness centroid $\mathbf{E_t}$ for every decade $t \in T = \{1850, 1860, \ldots, 1990\}$. As in Section 4, the extremeness centroid $\mathbf{E_t}$ is calculated as the averaged diachronic embeddings of the extreme adjectives introduced in Section 3. This provides us with a measure of EXTREMENESS for each decade $t \in T$ that we then average to get a single measure of EXTREMENESS for q.

SIMADJMOD. We calculate the odds-weighted[10] average similarity of the adverb embedding $\mathbf{q_t}$ to the embeddings of the adjectives $\mathbf{a_t}$, where a was modified by q in decade t. We perform this computation for every $t \in T = \{1850, 1860, \ldots, 1990\}$, giving us a measure of SIMADJMOD for every

[10] See Luo et al. (2019) for details of the weighting function.

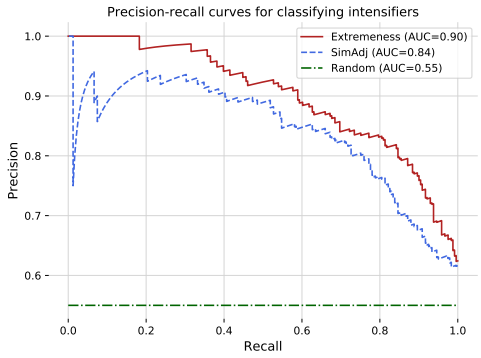

Figure 5: Results for classifying an adverb as an intensifier vs. control.

decade $t \in T$. We then average these per-decade SIMADJMOD values to get a single measure of SIMADJMOD for q. Rather than using the embedding of the *adjectival basis* of q for $\mathbf{q_t}$, as we did in Section 4 for data sparsity reasons, here we used the diachronic embeddings of the *adverb* q. In doing so, we mirror the operationalization of Luo et al. (2019) for this feature exactly.

B.3 Classification results

We fit a separate logistic regression classifier for each of the two features and use leave-one-out cross-validation. Fig. 5 shows that EXTREMENESS achieves better performance (AUC $= 0.90$) than SIMADJMOD (AUC $= 0.84$). The difference between the two features reported here is, however, much smaller than in the experiment in Section 4. Interestingly, while the AUC for EXTREMENESS is comparable between the two experiments, the SIMADJMOD feature does much better in the current set-up. We speculate that this is due to the latter feature having access to the adverbial vectors for decades in which the adverb has already obtained an intensifying meaning: a more "bleached" vector is more likely to be similar to many adjectives than a vector that has a more unique contextual signature. Further research is required on the influence of semantic bleaching on semantic space representations to test this speculation. We leave this topic for future work.

Despite the smaller difference in performance, however, we still find salient differences between the predictions made by the two models: We observe a similar pattern of errors as in Section 4 when setting a classification threshold at $P(c = \text{Intensifier}|q) = 0.5$. Specifically, the EXTREMENESS feature is able to correctly classify intensifiers that are relatively infrequent (*defiantly, startlingly, unequivocally*), providing further evidence that extremeness leads to the acquisition of intensifying meaning. Again, we observe that the account formalized in the SIMADJMOD measure correctly predicts adverbs with an epistemic function (*evidently, prominently, noticeably*) to acquire an intensifying meaning while the EXTREMENESS feature does not. Overall, we find results in accordance with those of Section 4: the cross-domain semantic property of EXTREMENESS is very effective in predicting whether an adverb will gain an intensifying function.