A Computational Framework for Slang Generation

Zhewei Sun[1], Richard Zemel[1,2], Yang Xu[1,2]

[1]Department of Computer Science, University of Toronto, Toronto, Canada
[2]Vector Institute for Artificial Intelligence, Toronto, Canada
`{zheweisun, zemel, yangxu}@cs.toronto.edu`

Abstract

Slang is a common type of informal language, but its flexible nature and paucity of data resources present challenges for existing natural language systems. We take an initial step toward machine generation of slang by developing a framework that models the speaker's word choice in slang context. Our framework encodes novel slang meaning by relating the conventional and slang senses of a word while incorporating syntactic and contextual knowledge in slang usage. We construct the framework using a combination of probabilistic inference and neural contrastive learning. We perform rigorous evaluations on three slang dictionaries and show that our approach not only outperforms state-of-the-art language models, but also better predicts the historical emergence of slang word usages from 1960s to 2000s. We interpret the proposed models and find that the contrastively learned semantic space is sensitive to the similarities between slang and conventional senses of words. Our work creates opportunities for the automated generation and interpretation of informal language.

1 Introduction

Slang is a common type of informal language that appears frequently in daily conversations, social media, and mobile platforms. The flexible and ephemeral nature of slang (Eble, 1989; Landau, 1984) poses a fundamental challenge for computational representation of slang in natural language systems. As of today, slang constitutes only a small portion of text corpora used in the natural language processing (NLP) community, and it is severely under-represented in standard lexical resources (Michel et al., 2011). Here we propose a novel framework for automated generation of slang with a focus on generative modeling of slang word meaning and choice.

Existing language models trained on large-scale text corpora have shown success in a variety of NLP tasks. However, they are typically biased toward formal language and under-represent slang. Consider the sentence ''I have a feeling he's gonna __ himself someday''. Directly applying a state-of-the-art GPT-2 (Radford et al., 2019) based language infilling model (e.g., Donahue et al., 2020) would result in the retrieval of *kill* as the most probable word choice (probability = 7.7%). However, such a language model is limited and near-insensitive to slang usage, for example, *ice*—a common slang alternative for *kill*—received virtually 0 probability, suggesting that existing models of distributional semantics, even the transformer-type models, do not capture slang effectively, if at all.

Our goal is to extend the capacity of NLP systems toward slang in a principled framework. As an initial step, we focus on modeling the generative process of slang, specifically the problem of slang word choice that we illustrate in Figure 1. Given an intended slang sense such as ''to kill'', we ask how we can emulate the speaker's choice of slang word(s) in informal context.[1] We are particularly interested in how the speaker chooses existing words from the lexicon and makes innovative use of those words in novel slang context (such as the use of *ice* in Figure 1).

Our basic premise is that sensible slang word choice depends on linking conventional or established senses of a word (such as ''frozen water'' for *ice*) to its emergent slang senses (such as ''to kill'' for *ice*). For instance, the extended use of *ice* to express killing could have emerged from the use of cold ice to refrigerate one's remains. A principled semantic representation should adapt to such similarity relations. Our proposed framework is aimed at encoding slang that relates informal

---

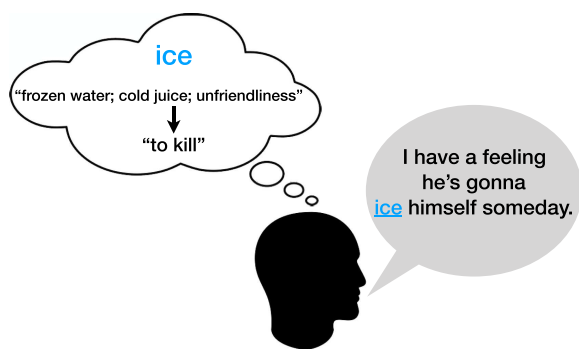[1]We will use the terms *meaning* and *sense* interchangeably.

Figure 1: A slang generation framework that models speaker's choice of a slang term (*ice*) based on the novel sense (''to kill'') in context and relations with conventional senses (e.g., ''frozen water'').

and conventional word senses, hence capturing semantic similarities beyond those from existing language models. In particular, contextualized embedding models such as BERT would consider ''frozen water'' to be semantically distant or irrelevant from ''to kill'', so they cannot predict *ice* to be appropriate for expressing ''to kill'' in slang context.

The capacity for generating novel slang word usages will have several implications and applications. From a scientific view, modeling the generative process of slang word choice will help explain the emergence of novel slang usages over time—we show how our framework can predict the emergence of slang in the history of English. From a practical perspective, automated slang generation paves the way for automated slang interpretation. Existing psycho-linguistic work suggests that language generation and comprehension rely on similar cognitive processes (e.g., Pickering and Garrod, 2013; Ferreira Pinto Jr. and Xu, 2021). Similarly, a generative model of slang can be an integral component of slang comprehension that informs the relation between a candidate sense and a query word, where the mapping can be unseen during training. Furthermore, a generative approach to slang may also be applied to downstream tasks such as naturalistic chatbots, sentiment analysis, and sarcasm detection (see work by Aly and van der Haar [2020] and Wilson et al., [2020]).

We propose a neural-probabilistic framework that involves three components: 1) a probabilistic choice model that infers an appropriate word for expressing a query slang meaning given its context, 2) an encoder based on contrastive learning that captures slang meaning in a modified embedding space, and 3) a prior that incorporates different forms of context. Specifically, the slang encoder we propose transforms slang and conventional senses of a word into a slang-sensitive embedding space where they will lie in close proximity. As such, senses like ''frozen water'', ''unfriendliness'' and ''to kill'' will be encouraged to be in close proximity in the learned embedding space. Furthermore, the resulting embedding space will also set apart slang senses of a word from senses of other unrelated words, and hence contrasting within-word senses from across-word senses in the lexicon. A practical advantage of this encoding method is that semantic similarities pertinent to slang can be extracted automatically from a small amount of training data, and the learned semantic space will be sensitive to slang.

Our framework also captures the flexible nature of slang usages in natural context. Here, we focus on syntax and linguistic context, although our framework should allow for the incorporation of social or extra-linguistic features as well. Recent work has found that the flexibility of slang is reflected prominently in syntactic shift (Pei et al., 2019). For example, *ice*—most commonly used as a noun—is used as a verb to express ''to kill'' (in Figure 1). We build on these findings by incorporating syntactic shift as a prior in the probabilistic model, which is integrated coherently with the contrastive neural encoder that captures flexibility in slang sense extension. We also show how a contextualized language infilling model can provide additional prior information from linguistic context (c.f. Erk, 2016).

To preview our results, we show that our framework yields a substantial improvement on the accuracy of slang generation against state-of-the-art embedding methods including deep contextualized models, in both few-shot and zero-shot settings. We evaluate our framework rigorously on three datasets constructed from slang dictionaries and in a historical prediction task. We show evidence that the learned slang embedding space yields intuitive interpretation of slang and offers future opportunities for informal natural language processing.

2 Related Work

NLP for Non-Literal Language. Machine processing of non-literal language has been explored in different linguistic phenomena including metaphor (Shutova et al., 2013b; Veale et al., 2016; Gao et al., 2018; Dankers et al., 2019), metonymy (Lapata and Lascarides, 2003; Nissim and Markert, 2003; Shutova et al., 2013a), irony (Filatova, 2012), neologism (Cook, 2010), idiom (Fazly et al., 2009; Liu and Hwa, 2018), vulgarity (Holgate et al., 2018), and euphemism (Magu and Luo, 2018). Non-literal usages are present in slang, but these existing studies do not directly model the semantics of slang. In addition, work in this area has typically focused on detection and comprehension. In contrast, generation of novel informal language use has been sparsely tackled.

Computational Studies of Slang. Slang has been extensively studied as a social phenomenon (Mattiello, 2005), where social variables such as gender (Blodgett et al., 2016), ethnicity (Bamman et al., 2014), and social-economic status (Labov, 1972, 2006) have been shown to play important roles in slang construction. More recently, an analysis of social media text has shown that linguistic features also correlate with the survival of slang terms, where linguistically appropriate terms have a higher likelihood of being popularized (Stewart and Eisenstein, 2018).

Recent work in the NLP community has also analyzed slang. Ni and Wang (2017) studied slang comprehension as a translation task. In their study, both the spelling of a word and its context are provided as input to a translation model to decode a definition sentence. Pei et al. (2019) proposed end-to-end neural models to detect and identify slang automatically in natural sentences.

Kulkarni and Wang (2018) have proposed computational models that derive novel word forms of slang from spellings of existing words. Here, we instead explore the generation of novel slang usage from existing words and focus on word sense extension toward slang context, based on the premise that new slang senses are often derived from existing conventional word senses.

Our work is inspired by previous research suggesting that slang relies on reusing words in the existing lexicon (Eble, 2012). Previous work has applied cognitive models of categorization to predict novel slang usage (Sun et al., 2019). In that work, the generative model is motivated by research on word sense extension (Ramiro et al., 2018). In particular, slang generation is operationalized by categorizing slang senses based on their similarities to dictionary definitions of candidate words, enhanced by collaborative filtering (Goldberg et al., 1992) to capture the fact that words with similar senses are likely to extend to similar novel senses (Lehrer, 1985). However, this approach presupposes that slang senses are similar to conventional senses of a word represented in standard embedding space, an assumption that is not warranted and yet to be addressed.

Our work goes beyond the existing work in three important aspects: 1) We capture semantic flexibility of slang usage by contributing a novel method based on contrastive learning. Our method encodes slang meaning and conventional meaning of a word under a common embedding space, thereby improving the inadequate existing methodology for slang generation that uses common, slang-insensitive embeddings. 2) We capture syntactic and contextual flexibility of slang usage in a coherent probabilistic framework, an aspect that was ignored in previous work. 3) We rigorously test our framework against slang sense definition entries from three large slang dictionaries and contribute a new dataset for slang research to the community.

Contrastive Learning. Contrastive learning is a semi-supervised learning technique used to extract semantic representations in data-scarce situations. It can be incorporated into neural networks in the form of twin networks, where two exact copies of an encoder network are applied to two different examples. The encoded representations are then compared and back-propagated. Alternative loss schemes such as Triplet (Weinberger and Saul, 2009; Wang et al., 2014) and Quadruplet loss (Law et al., 2013) have also been proposed to enhance stability in training. In NLP, contrastive learning has been applied to learn similarities between text (Mueller and Thyagarajan, 2016; Neculoiu et al., 2016) and speech utterances (Kamper et al., 2016) with recurrent neural networks.

The contrastive learning method we develop has two main differences: 1) We do not use recurrent encoders because they perform poorly on dictionary-based definitions; 2) We propose a joint neural-probabilistic framework on the learned embedding space instead of resorting to methods such as nearest-neighbor search for generation.

3 Computational Framework

Our computational framework for slang generation comprises three interrelated components: 1) A probabilistic formulation of word choice, extending that in Sun et al. (2019) to leverage encapsulated slang senses from a modified embedding space; 2) A contrastive encoder—inspired by variants of twin network (Baldi and Chauvin, 1993; Bromley et al., 1994)—that constructs a modified embedding space for slang by adapting the conventional embeddings to incorporate new senses of slang words; 3) A contextually informed prior for capturing flexible uses of naturalistic slang.

3.1 Probabilistic Slang Choice Model

Given a query slang sense M_S and its context C_S, we cast the problem of slang generation as inference over candidate words w in our vocabulary. Assuming all candidate words w are drawn from a fixed vocabulary V, the posterior is as follows:

$$P(w|M_S, C_S) \propto P(M_S|w, C_S)P(w|C_S)$$
$$\propto P(M_S|w)P(w|C_S) \quad (1)$$

Here, we define the prior $P(w|C_S)$ based on regularities of syntax and/or linguistic context in slang usage (described in Section 3.4). We formulate the likelihood $P(M_S|w)^2$ by specifying the relations between conventional senses of word w (denoted by $\mathcal{M}_w = \{M_{w_1}, M_{w_2}, \cdots, M_{w_m}\}$, i.e., the set of senses drawn from a standard dictionary) and the query M_S (i.e., slang sense that is outside the standard dictionary). Specifically, we model the likelihood by a similarity function that measures the proximity between the slang sense M_S and the set of conventional senses $\mathcal{M}_w$ of word w in a continuous, embedded semantic space:

$$P(M_S|w) = P(M_S|\mathcal{M}_w)$$
$$\propto f(\{sim(E_S, E_{w_i}); E_{w_i} \in \mathcal{E}_w\}) \quad (2)$$

Here, $f(\cdot)$ is a similarity function in range $[0, 1]$, while E_S and $\mathcal{E}_w$ represent semantic embeddings of the slang sense M_S and the set of conventional senses $\mathcal{M}_w$. We derive these embeddings from contrastive learning which we describe in detail in Section 3.2, and we compare this proposed method

---

²Here, we only consider linguistically motivated context as C_S and assume the semantic shift patterns of slang are universal across all such contexts.

with baseline methods that draw embeddings from existing sentence embedding models.

Our choice of the similarity function $f(\cdot)$ is motivated by prior work on few-shot classification. Specifically, we consider variants of two established methods: One Nearest Neighbor (1NN) matching (Koch et al., 2015; Vinyals et al., 2016) and Prototypical learning (Snell et al., 2017).

The 1NN model postulates that a candidate word should be chosen according to the similarity between the query slang sense and the closest conventional sense:

$$f_{1nn}(E_S, \mathcal{E}_w) = \max_{E_{w_i} \in \mathcal{E}_w} sim(E_S, E_{w_i}) \quad (3)$$

In contrast, the prototype model postulates that a candidate word should be chosen if its aggregate (or average) sense is in close proximity of the query slang sense:

$$f_{prototype}(E_S, \mathcal{E}_w) = sim(E_S, E_w^{prototype})$$
$$E_w^{prototype} = \frac{1}{|\mathcal{E}_w|} \sum_{E_{w_i} \in \mathcal{E}_w} E_{w_i} \quad (4)$$

In both cases, the similarity between two senses is defined by the exponentiated negative squared Euclidean distance in semantic embedding space:

$$sim(E_S, E_w) = \exp(-\frac{||E_S - E_w||_2^2}{h_s}) \quad (5)$$

Here, h_s is a learned kernel width parameter.

We also consider an enhanced version of the posterior using collaborative filtering (Goldberg et al., 1992), where words with similar meaning are predicted to shift to similar novel slang meanings. We operationalize this by summing over the close neighborhood of candidate word $L(w)$:

$$P(w|M_S, C_S) = \sum_{w' \in L(w)} P(w|w')P(w'|M_S, C_S) \quad (6)$$

Here, $P(w'|M_S, C_S)$ is a fixed term calculated identically as in Equation (1) and $P(w|w')$ is the weighting of words in the close neighborhood of a candidate word w. This weighting probability is set proportional to the exponentiated negative cosine distance between w and its neighbor w' defined in pre-trained word embedding space, and the kernel parameter h_{cf} is also estimated from the training data:

$$P(w|w') \propto sim(w, w') = \exp(-\frac{d(w, w')}{h_{cf}}) \quad (7)$$

Here, $d(w, w')$ is the cosine distance between two words in a word embedding space.

3.2 Contrastive Semantic Encoding (CSE)

We develop a contrastive semantic encoder for constructing a new embedding space representing slang and conventional word senses that do not bear surface similarities. For instance, the conventional sense of *kick* such as ''propel with foot'' can hardly be related to the slang sense of *kick* such as ''a strong flavor''. The contrastive embedding space we construct seeks to redefine or warp similarities, such that the otherwise unrelated senses will be in closer proximity than they are under existing embedding methods. For example, two metaphorically related senses can bear strong similarity in slang usage, even though they may be far apart in a literal sense.

We sample triplets of word senses as input to contrastive learning, following work on twin networks (Baldi and Chauvin, 1993; Bromley et al., 1994; Chopra et al., 2005; Koch et al., 2015). We use dictionary definitions of conventional and slang senses to obtain the initial sense embeddings (See Section 4.4 for details). Each triplet consists of 1) an anchor slang sense M_S, 2) a positive conventional sense M_P, and 3) a negative conventional sense M_N. The positive sense should ideally be encouraged to lie closely to the anchor slang sense (in the resulting embedding space), whereas the negative sense should ideally be further away from both the positive conventional and anchor slang senses. Section 3.3 describes the detailed sampling procedures.

Our triplet network uses a single neural encoder g to project each word sense representation into a joint embedding space in $\mathbb{R}^d$.

$$E_S = g(M_S); E_P = g(M_P); E_N = g(M_N) \tag{8}$$

We choose a 1-layer fully connected network with ReLU(Nair and Hinton, 2010) as the encoder g for pre-trained word vectors (e.g., fastText). For contextualized embedding models we consider, g will be a Transformer encoder (Vaswani et al., 2017). In both cases, we apply the same encoder network g to each of the three inputs. We train the triplet network using the max-margin triplet loss (Weinberger and Saul, 2009), where the squared distance between the positive pair is constrained

to be closer than that of the negative pair with a margin m:

$$L_{triplet} = \left[m + \|E_S - E_P\|_2^2 - \|E_S - E_N\|_2^2 \right]_+ \tag{9}$$

3.3 Triplet Sampling

To train the triplet network, we build data triplets from every slang lexical entry in our training set. For each slang sense M_S of word w, we create a positive pair with each conventional sense M_{w_i} of the same word w. Then for each positive pair, we sample a negative example every training epoch by randomly selecting a conventional sense $M_{w'}$ from a word w' that is sufficiently different from w, such that the corresponding definition sentence $D_{w'}$ has less than 20% overlap in the set of content words compared to M_S and any conventional definition sentence D_{w_i} of word w. We rank all candidate words in our vocabulary against w by computing cosine distances from pre-trained word embeddings and consider a word w' to be sufficiently different if it is not in the top 20 percent.

Neighborhood Sampling (NS). In addition to using conventional senses of the matching word w for constructing positive pairs, we also sample positive senses from a small neighborhood $L(w)$ of similar words. This sampling strategy provides linguistic knowledge from parallel semantic change to encourage neighborhood structure in the learned embedding space. Sampling from neighboring words also augments the size of the training data considerably in this data-scarce task. We sample negative senses in a similar way, except that we also consider all conventional definition sentences from neighboring words when checking for overlapping senses.

3.4 Contextual Prior

The final component of our framework is the prior $P(w|C_S)$ (see Equation (1)) that captures flexible use of slang words with regard to syntax and distributional semantics. For example, slang exhibits flexible Part-of-Speech (POS) shift, for example, noun→verb transition as in the example *ice*, and surprisals in linguistic context, for example, *ice* in ''I have a feeling he's gonna [blank] himself someday.'' Here, we formulate the context C_S in two forms: 1) a syntactic-shift prior, namely, the POS

information P_S to capture syntactic regularities in slang, and/or 2) a linguistic context prior, namely, the linguistic context K_S to capture distributional semantic context when this is available in the data.

Syntactic-Shift Prior (SSP). Given a query POS tag P_S, we construct the syntactic prior by comparing POS distribution $\mathcal{P}_w$ from literal natural usage of a candidate word w with a smoothed POS distribution $\mathcal{P}_S$ centered at P_S. However, we cannot directly compare $\mathcal{P}_S$ to $\mathcal{P}_w$ because slang usage often involves shifting POS (Eble, 2012; Pei et al., 2019). To account for this, we apply a transformation T by counting the number of POS transitions for each slang-conventional definition pair in the training data. Each column of the transformation matrix T is then normalized, so column i of T can be interpreted as the expected slang-informed POS distribution given the i'th POS tag in conventional context (e.g., the noun column gives the expected slang POS distribution of a word that is used exclusively as a noun in conventional usage). The slang-contextualized POS distribution $\mathcal{P}_S^*$ can then be computed by applying T on $\mathcal{P}_S$: $\mathcal{P}_S^* = T \times \mathcal{P}_S$. The prior can be estimated by comparing the POS distributions $\mathcal{P}_w$ and $\mathcal{P}_S^*$ via Kullback-Leibler (KL) divergence:

$$P(w|C_S) = P(w|P_S) \propto \exp\left(-KL(\mathcal{P}_w, \mathcal{P}_S^*)\right)^{\frac{1}{2}}$$
(10)

Intuitively, this prior captures the regularities of syntactic shift in slang usage, and it favors candidate words with POS characteristics that fits well with the queried POS tag in a slang context.

Linguistic Context Prior (LCP). We use a language model P_{LM} to a given linguistic context K_S to estimate the probability of each candidate word:

$$P(w|C_S) = P(w|K_S) \propto P_{LM}(w|K_S) + \alpha \quad (11)$$

Here, α is a Laplace smoothing constant. We use the GPT-2 based language infilling model from Donahue et al. (2020) as P_{LM} and discuss the implementation in Section 4.3.

4 Experimental Setup

4.1 Lexical Resources

We collected lexical entries of slang and conventional words/phrases from three separate online dictionaries:[3] 1) Online Slang Dictionary (OSD),[4] 2) Green's Dictionary of Slang (GDoS) (Green, 2010),[5] and 3) an open source subset of Urban Dictionary (UD) data from Kaggle.[6] In addition, we gathered dictionary definitions of conventional senses of words from the online version of Oxford Dictionary (OD).[7]

Slang Dictionary. Both slang dictionaries (OSD and GDoS) are freely accessible online and contain slang definitions with meta-data such as Part-of-Speech tags. Each data entry contains the word, its slang definition, and its part-of-speech (POS) tag. In particular, OSD includes example sentence(s) for each slang entry which we leverage as linguistic context, and GDoS contains time-tagged references that allow us to perform historical prediction (described later). We removed all acronyms (i.e., fully capitalized words) as they generally do not extend meaning, and slang definitions that share more than 50% content words with any of their corresponding conventional definitions to account for conventionalized slang. We also removed slang with novel word forms where no conventional sense definitions are available. Slang phrases were treated as unigrams because our task only concerns the association between senses and lexical items. Each sense definition was considered a data point during both learning and prediction. We later partitioned definition entries from each dataset to be used for training, validation, and testing. Note that a word may appear in both training and testing but the pairing between word senses are unique (See Section 5.3 for discussion).

Conventional Word Senses. We focused on the subset of OD containing word forms that are also available in the slang datasets described. For each word entry, we removed all definitions that have been tagged as *informal* because these are likely to represent slang senses. This results in 10,091 and 29,640 conventional sense definitions corresponding to the OSD and GDoS datasets, respectively.

Data Split. We used all definition entries from the slang resources such that the corresponding slang word also exists in the collected OD subset. The resulting datasets (OSD and GDS) had 2,979 and 29,300 definition entries, respectively, from 1,635 and 6,540 unique slang words, of which 1,253 are shared across both dictionaries. For each dataset, the slang definition entries were partitioned into a 90% training set and a 10% test set. Five percent of the data in the training set were set aside for validation when training the contrastive encoder.

Urban Dictionary. In addition to the two datasets described above, we provide a third dataset based on Urban Dictionary (UD) that are made available via Kaggle. Unlike the previous two datasets, we are able to make this one publicly available without requiring one to obtain prior permission from the data owners.[8] To guard against the crowd-sourced and noisy nature of UD, we ensure quality by keeping definition entries such that 1) it has at least 10 more upvotes than downvotes, 2) the word entry exists in one of OSD or GDoS, and 3) at least one of the corresponding definition sentences in these dictionaries have a 20% or greater overlap in the set of content words with the UD definition sentence. We also remove entries with more than 50% overlap in content words with any other UD slang definitions under the same word to remove duplicated senses. This results in 2,631 definitions entries from 1,464 unique slang words. The corresponding OD subset has 10,357 conventional sense entries. We find entries from UD to be more stylistically variable and lengthier, with a mean entry length of 9.73 in comparison to 7.54 and 6.48 for OSD and GDoS, respectively.

4.2 Part-of-Speech Data

The natural POS distribution $\mathcal{P}_w$ for each candidate word w is obtained using POS counts from the most recent available decade of the HistWords project (Hamilton et al., 2016). For word entries that are not available, mostly phrases, we estimate $\mathcal{P}_w$ by counting POS tags from Oxford Dictionary (OD) entries of w.

When estimating the slang POS transformation for the syntactic prior, we mapped all POS tags into one of the following six categories: {verb,

---

other, adv, noun, interj, adj} for the OSD experiments. For GDS, the tag 'interj' was excluded as it is not present in the dataset.

4.3 Contextualized Language Model Baseline

We considered a state-of-the-art GPT-2 based language infilling model from Donahue et al. (2020) as both a baseline model and a prior to our framework (on the OSD data where context sentences are available for the slang entries). For each entry, we blanked out the corresponding slang word in the example sentence, effectively treating our task as a cloze task. We applied the infilling model to obtain probability scores for each of the candidate words and apply a Laplace smoothing of 0.001. We fine-tuned the LM infilling model using all example sentences in the OSD training set until convergence. We also experiment with a combined prior where the two priors are combined using element-wise multiplication and re-normalization.

4.4 Baseline Embedding Methods

To compare with and compute the baseline embedding methods M for definition sentences, we used 300-dimensional fastText embeddings (Bojanowski et al., 2017) pre-trained with sub-word information on 600 billion tokens from Common Crawl[9] as well as 768-dimensional Sentence-Bert (SBERT) (Reimers and Gurevych, 2019) encoders pretrained on Wikipedia and fine-tuned on NLI datasets (Bowman et al., 2015; Williams et al., 2018). The fastText embeddings were also used to compute cosine distances $d(w, w')$ in Equation (7). Embeddings for phrases and the fastText-based sentence embeddings were both computed by applying average pooling to normalized word-level embeddings of all content words. In the case of SBERT, we fed in the original definition sentence.

4.5 Training Procedures

We trained the triplet networks for a maximum of 20 epochs using Adam (Kingma and Ba, 2015) with a learning rate of 10^{-4} for fastText and 2^{-5} for SBERT based models. We preserved dimensions of the input sense vectors for the contrastive embeddings learned by the triplet network (that is, 300 for fastText and 768 for SBERT). We used 1,000 fully-connected units in the contrastive

---

encoder's hidden layer for fastText based models. Triplet margins of 0.1 and 1.0 were used with fastText and SBERT embeddings respectively.

We trained the probabilistic classification framework by minimizing negative log likelihood of the posterior $P(w^*|M_S, C_S)$ on the ground-truth words for all definition entries in the training set. We jointly optimized kernel width parameters using L-BFGS-B (Byrd et al., 1995). To construct a word w's neighborhood $L(w)$ in both collaborative filtering and triplet sampling, we considered the 5 closest words in cosine distances of their fastText embeddings.

5 Results

5.1 Model Evaluation

We first evaluated our models quantitatively by predicting slang word choices: Given a novel slang sense (a definition taken from a slang dictionary) and its part-of-speech, how likely is the model to predict the ground-truth slang recorded in the dictionary? To assess model performance, we allowed each model to make up to $|V|$ ranked predictions where V is the vocabulary of the dataset being evaluated, and we used standard Area-Under-Curve (AUC) percentage from Receiver-Operator Characteristic (ROC) curves to assess overall performance.

We show the ROC curves for the OSD evaluation in Figure 2 as an illustration. The AUC metric is similar to and a continuous extension to an F1 score by comprehensively sweeping through the number of candidate words a model is allowed to predict. We find this metric to be the most appropriate because multiple words may be appropriate to express a probe slang sense.

To examine the effectiveness of the contrastive embedding method, we varied the semantic representation as input to the models by considering both fastText and SBERT (described in Sec 4.4). For both embeddings, we experimented with the baseline variant without the contrastive encoding (e.g., vanilla embeddings from fastText and SBERT). We then augmented the models incrementally with the contrastive encoder and the priors whenever applicable to examine their respective and joint effects on model performance in slang word choice prediction. We observed that, under both datasets, models leveraging the contrastively learned sense embeddings more reliably predict the ground-truth slang words, indicated by
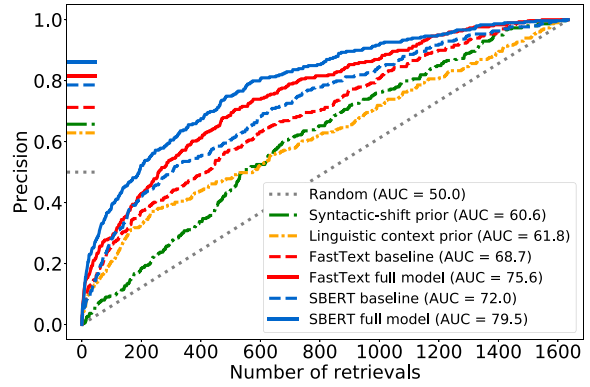

Figure 2: ROC curves for slang generation in OSD test set. Collaborative-filtering prototype model was used for prediction. Ticks on the y-axis indicate median precision of the models.

both higher AUC scores and consistent improvement in precision over all retrieval ranks. Note that the vanilla SBERT model, despite being a much larger model trained on more data, only presented minor performance gains when compared with the plain fastText model. This suggests that simply training larger models on more data does not better encapsulate slang semantics.

We also analyzed whether the contrastive embeddings are robust under different choices of the probabilistic models. Specifically, we considered the following four variants of the models: 1) 1-Nearest Neighbor (1NN), 2) Prototype, 3) 1NN with collaborative filtering (CF), and 4) Prototype with CF. Our results show that applying contrastively learned semantic embeddings consistently improves predictive accuracy across all probabilistic choice models. The complete set of results for all 3 datasets is summarized in Table 1.

We noted that the syntactic information from the prior improves predictive accuracy in all settings, while by itself predicting significantly better than chance. On OSD, we used the context sentences alone in a contextualized language infilling model for prediction and also incorporating it as a prior. Again, while the prior consistently improves model prediction, both by itself and when paired with the syntactic-shift prior, the language model alone is not sufficient.

We found the syntactic-shift prior and linguistic context prior to be capturing complementary information (mean Spearman correlation of 0.054 ± 0.003 across all examples), resulting in improved performance when they are combined together.

Model	1NN	Prototype	1NN+CF	Proto+CF
Dataset 1: Online Slang Dictionary (OSD)				
Prior Baseline - Uniform			51.9	
Prior Baseline - Syntactic-shift			60.6	
Prior Baseline - Linguistic Context(Donahue et al., 2020)			61.8	
Prior Baseline - Syntactic-shift + Linguistic Context			**67.3**	
FastText Baseline(Sun et al., 2019)	63.2	65.2	66.0	68.7
FastText + Contrastive Semantic Encoding (CSE)	71.7	71.6	73.0	72.6
FastText + CSE + Syntactic-shift Prior (SSP)	73.8	73.4	75.2	74.4
FastText + CSE + Linguistic Context Prior (LCP)	73.6	73.2	74.7	73.9
FastText + CSE + SSP + LCP	**75.4**	**74.9**	**76.5**	**75.6**
SBERT Baseline	67.4	68.1	69.5	72.0
SBERT + CSE	76.6	77.4	77.4	78.0
SBERT + CSE + SSP	77.6	78.0	78.8	78.9
SBERT + CSE + LCP	77.8	78.4	78.1	78.7
SBERT + CSE + SSP + LCP	**78.5**	**79.0**	**79.4**	**79.5**
Dataset 2: Green's Dictionary of Slang (GDoS)				
Prior Baseline - Uniform			51.5	
Prior Baseline - Syntactic-shift			**61.0**	
FastText Baseline(Sun et al., 2019)	68.2	69.9	67.8	69.7
FastText + Contrastive Semantic Encoding (CSE)	73.4	74.0	74.1	74.8
FastText + CSE + Syntactic-shift Prior (SSP)	**74.5**	**74.8**	**75.2**	**75.8**
SBERT Baseline	67.1	68.0	66.8	67.5
SBERT + CSE	77.8	78.2	77.4	77.9
SBERT + CSE + SSP	**78.5**	**78.7**	**78.3**	**78.6**
Dataset 3: Urban Dictionary (UD)				
Prior Baseline - Uniform			52.3	
FastText Baseline(Sun et al., 2019)	65.2	68.8	67.6	70.9
FastText + Contrastive Semantic Encoding (CSE)	**71.0**	**72.2**	**71.5**	**73.7**
SBERT Baseline	72.4	71.7	74.0	74.4
SBERT + CSE	**76.2**	**76.6**	**77.2**	**78.8**

Table 1: Summary of model AUC scores (%) for slang generation in 3 slang datasets.

However, the majority of the performance gain is attributed to the augmented contrastive embeddings, which highlights the importance and supports our premise that encoding of slang and conventional senses is crucial to slang word choice.

5.2 Historical Analysis of Slang Emergence

We next performed a temporal analysis to evaluate whether our model explains slang emergence over time. We used the time tags available in the GDoS dataset and predicted historically emerged slang from the past 50 years (1960s–2000s). For a given slang entry recorded in history, we tagged its emergent decade using the earliest dated reference available in the dictionary. For each future decade d, we trained our model using all entries before

d and assessed whether our model can predict the choices of slang words for slang senses that emerged in the future decade. We scored the models on slang words that emerged during each subsequent decade, simulating a scenario where future slang usages are incrementally predicted.

Table 2 summarizes the result from the historical analysis for the non-contrastive SBERT baseline and our full model (with contrastive embeddings), based on the GDoS data. AUC scores are similar to the previous findings but slightly lower for both models in this historical setting. Overall, we find the full model to improve the baseline consistently over the course of history examined and achieve similar performance as in the synchronic evaluation. This provides strong evidence that our framework is robust and has

Decade	# Test	Baseline	SBERT+CSE+SSP
1960s	2010	67.5	**77.4**
1970s	1757	66.3	**77.9**
1980s	1655	66.3	**78.6**
1990s	1605	66.2	**75.4**
2000s	1374	65.9	**77.0**

Table 2: Summary of model AUC scores in historical prediction of slang emergence (1960s-2000s). The non-contrastive SBERT baseline and the proposed full model (with contrastive embedding, CSE, and syntactic prior, SSP) are compared using collaborative-filtering Prototype. Models were trained and tested incrementally through time (test sizes shown) and trained initially on 20,899 Green's Dictionary definitions prior to the 1960s.

explanatory power over the historical emergence of slang.

5.3 Model Error Analysis and Interpretation

Few-shot vs Zero-shot Prediction. We analyze our model errors and note that one source of error stems from whether the probe slang word has appeared during training versus not. Here, each candidate word is treated as a class and each slang sense of a word seen in the training set is considered a 'shot'. In the few-shot case, although the slang sense in question was not observed in prediction, the model has some *a priori* knowledge about its target word and how it has been used in slang context (because a word may have multiple slang senses), thus allowing the model to generalize toward novel slang usage of that word. In the zero-shot case, the model needs to select a novel slang word (i.e., one that never appeared in training) and hence has no direct knowledge about how that word should be extended in a slang context. Such knowledge must be inferred indirectly, and in this case, from the conventional senses of the candidate words. The model can then infer how words with similar conventional senses might extend to slang context.

Table 3 outlines the AUC scores of the collaboratively filtered prototype models under few-shot and zero-shot settings. For each dataset, we partitioned the corresponding test set by whether the target word appears at least once within another definition entry in the training data. This results in 179, 2,661, and 165 few-shot definitions in

(a) Online Slang Dictionary (OSD)		
Model	Few-shot	Zero-shot
Prior - Uniform	55.1	47.1
Prior - Syntactic-shift	63.4	**56.4**
Prior - Linguistic Context	72.4	45.8
Prior - SSP + LCP	**74.7**	**56.4**
FT Baseline	68.3	69.2
FT + CSE	74.8	69.4
FT + CSE + SSP	76.8	**70.9**
FT + CSE + LCP	76.7	69.5
FT + CSE + SSP + LCP	**78.7**	**70.9**
SBERT Baseline	72.2	71.6
SBERT + CSE	78.3	77.5
SBERT + CSE + SSP	79.3	**78.3**
SBERT + CSE + LCP	79.8	77.1
SBERT + CSE + SSP + LCP	**80.7**	77.8

(b) Green's Dictionary of Slang (GDoS)		
Model	Few-shot	Zero-shot
Prior - Uniform	51.8	48.1
Prior - Syntactic-shift	**61.6**	**54.8**
FT Baseline	70.6	**61.3**
FT + CSE	76.3	59.2
FT + CSE + SSP	**77.3**	60.7
SBERT Baseline	68.3	59.6
SBERT + CSE	79.0	66.8
SBERT + CSE + SSP	**79.7**	**67.7**

(c) Urban Dictionary (UD)		
Model	Few-shot	Zero-shot
Prior - Uniform	54.2	49.1
FT Baseline	68.6	**75.0**
FT + CSE	**76.2**	69.4
SBERT Baseline	73.0	**76.8**
SBERT + CSE	**80.6**	75.6

Table 3: Model AUC scores (%) for Few-shot and Zero-shot test sets (''CSE'' for contrastive embedding, ''SSP'' for syntactic prior, ''LCP'' for contextual prior, and ''FT'' for fastText).
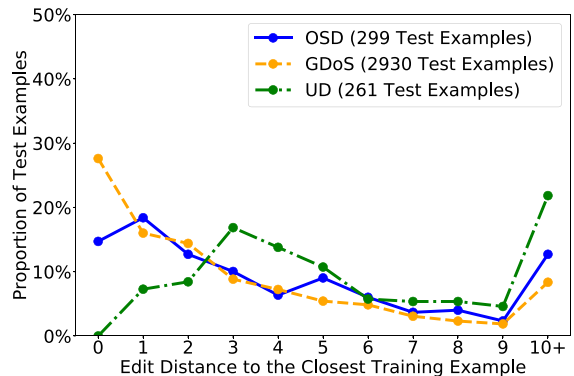

Figure 3: Degree of synonymy in the test examples relative to training data in each of the 3 datasets.

(a) Online Slang Dictionary (OSD)

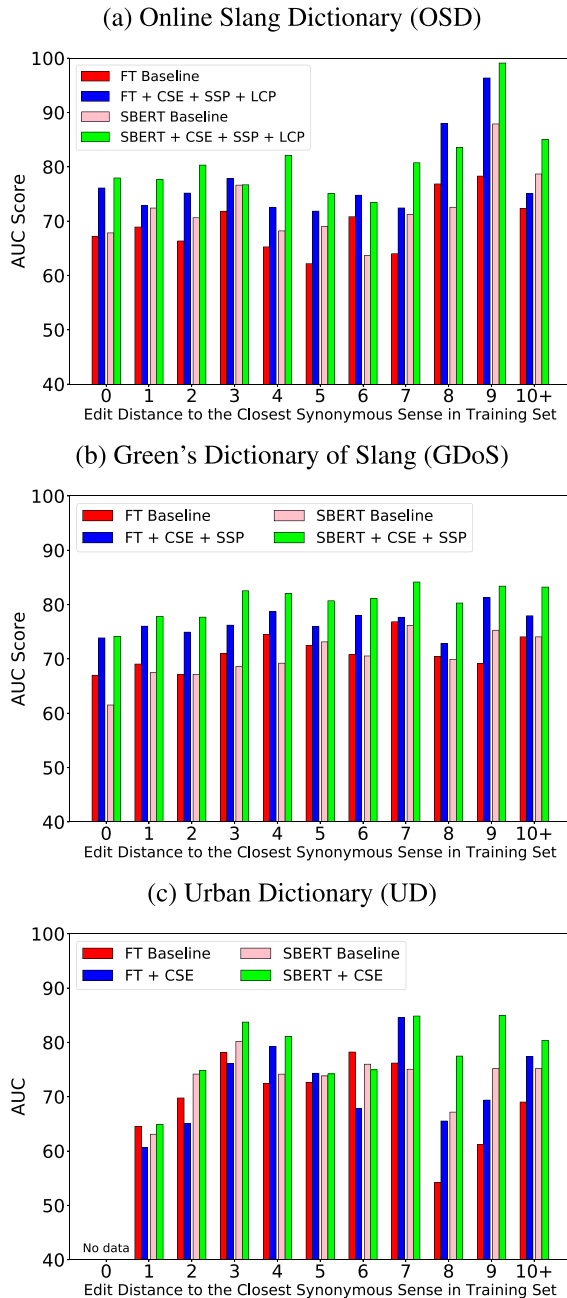(b) Green's Dictionary of Slang (GDoS)

(c) Urban Dictionary (UD)

Figure 4: Model AUC scores (%) under test sets with different degrees of synonymy present in training, for the baselines and the best performing models (under collaborative-filtering prototype).

OSD, GDoS, and UD, respectively, along with 120, 269, 96 zero-shot definitions. From our results, we observed that it is more challenging for the model to generalize usage patterns to unseen words, with AUC scores often being higher in the few-shot case. Overall, we found the model to have the most issues handling zero-shot cases from GDoS due to the fine-grained senses recorded in this dictionary, where a word has more slang senses on average (in comparison to

(a) Online Slang Dictionary (OSD)		
Model	Training	Testing
FT Baseline	0.33 ± 0.011	0.35 ± 0.033
FT + CSE	0.15 ± 0.0083	0.28 ± 0.030
SBERT Baseline	0.34 ± 0.011	0.32 ± 0.033
SBERT + CSE	0.097 ± 0.0069	0.23 ± 0.029
(b) Green's Dictionary of Slang (GDoS)		
Model	Training	Testing
FT Baseline	0.30 ± 0.0034	0.30 ± 0.010
FT + CSE	0.19 ± 0.0028	0.26 ± 0.0097
SBERT Baseline	0.32 ± 0.0035	0.32 ± 0.010
SBERT + CSE	0.10 ± 0.0019	0.22 ± 0.0089
(c) Urban Dictionary (UD)		
Model	Training	Testing
FT Baseline	0.34 ± 0.012	0.31 ± 0.037
FT + CSE	0.20 ± 0.010	0.28 ± 0.033
SBERT Baseline	0.34 ± 0.012	0.28 ± 0.034
SBERT + CSE	0.10 ± 0.0075	0.23 ± 0.031

Table 4: Mean Euclidean distance from slang senses to prototypical conventional senses.

the OSD and UD data). This issue caused the models to be more biased towards generalizing usage patterns from more commonly observed words. Finally, the SBERT-based models tend to be more robust towards unseen word-forms, potentially benefiting from their contextualized properties.

Synonymous Slang Senses. We also examined the influence of synonymy (or sense overlap) in the slang datasets. We quantified the degree of sense synonymy by checking each test sense against all training senses and computing the edit distance between the corresponding sets of constituent content words of the sense definitions.

Figure 3 shows the distribution of degree of synonymy across all test examples where the edit distance to the closest training example is considered. We perform our evaluation by binning based on the degree of synonymy and summarize the results in Figure 4. We do not observe any substantial changes in performance when controlling for the degree of synonymy, and in fact, the highly synonymous definitions appear to be more difficult (as opposed to easier) for the models. Overall, we find the models to yield consistent improvement across different degrees of synonymy, particularly with the SBERT based full model, which offers improvement in all cases.

472

Model	Top-5 slang words predicted by model	Predicted rank of the true slang
1. True slang: *kick*; Slang sense: **"a thrill, amusement or excitement"**		
Sample usage: I got a huge *kick* when things were close to out of hand.		
SBERT Baseline	*thrill, pleasure, frolic, yahoo, sparkle*	3496 / 6540
Full model	*twist, spin, trick, crank, punch*	96 / 6540
2. True slang: *whiff*; Slang sense: **"to kill, to murder, [play on SE, to blow away]"**		
Sample usage: The trouble is he wasn't alone when you *whiffed* him.		
SBERT Baseline	*suicide, homicide, murder, killing, rape*	2735 / 6540
Full model	*spill, swallow, blow, flare, dash*	296 / 6540
3. True slang: *chirp*; Slang sense: **"an act of informing, a betrayal"**		
Sample usage: Once we're sure there's no back-fire anywhere, the Sparrow will *chirp* his last chirp.		
SBERT Baseline	*dupe, sin, scam, humbug, hocus*	2431 / 6540
Full model	*chirp, squeal, squawk, fib, chat*	1 / 6540
4. True slang: *red*; Slang sense: **"a communist, a socialist or anyone considered to have left-wing leanings"**		
Sample usage: Why the hell would I bed a *red*?		
SBERT Baseline	*leveller, wildcat, mole, pawn, domino*	1744 / 6540
Full model	*orange, bluey, black and tan, violet, shadow*	164 / 6540
5. True slang: *team*; Slang sense: **"a gang of criminals"**		
Sample usage: And a little *team* to follow me - all wanted up the yard.		
SBERT Baseline	*gangster, hoodlum, thug, mob, gangsta*	826 / 6540
Full model	*brigade, mob, business, gang, school*	15 / 6540

Table 5: Example slang word predictions from the contrastively learned full model and SBERT baseline (with no contrastive embedding) on slang usage from the Green's Dictionary. Each example shows the true slang, the probe slang sense, a sample usage, the alternative slang words predicted by each model, and the predicted rank (colored bars indicate inverse rank) of the true slang from a lexicon of 6,540 words.

Semantic Distance. To understand the consequence of contrastive embedding, we examine the relative distance between conventional and slang senses of a word in embedding space and the extent to which the learned semantic relations might generalize. We measured the Euclidean distance between each slang embedding with the prototype sense vector of all candidate words, without applying the probabilistic choice models. Table 4 shows the ranks of the corresponding candidate words, averaged over all slang sense embeddings considered and normalized between 0 and 1. We observed that contrastive learning indeed brings closer slang and conventional senses (from the same word), as indicated by lower mean semantic distance after the embedding procedure is applied. Under both fastText and SBERT, we obtained significant improvement on both the OSD and GDoS test sets ($p < 0.001$). On UD, the improvement is significant for SBERT ($p = 0.018$) but marginal for fastText ($p = 0.087$).

Examples of Model Prediction. Table 5 shows 5 example slangs from the GDoS test set and the top words predicted by both the baseline SBERT model and the full SBERT-based model with contrastive learning. The full model exhibits a greater tendency to choose words that appear remotely related to the queried sense (e.g., *spill*, *swallow* for the act of killing), while the baseline model favors words that share only surface semantic similarity (e.g., retrieving *murder* and *homicide* directly). We found cases where the model extends meaning metaphorically (e.g., animal to action, in the case of *chirp*), euphemistically (e.g., *spill* and *swallow* for kill), and generalization of a concept (e.g., *brigade* and *mob* for gang), all of which are commonly attested in slang usage (Eble, 2012).

We found the full model to achieve better retrieval accuracy in cases where the queried slang undergoes a non-literal sense extension, whereas the baseline model is situated at retrieving candidate words with incremental or literal changes in meaning. We also noted many cases where the true slang word is difficult to predict without appropriate background knowledge. For instance, the full-model suggested words such as *orange* and *bluey* to mean "a communist" but could not pinpoint the color *red* without knowing its cultural association to communism. Finally, we observed that our model to perform generally worse when the target slang sense can hardly be related to conventional senses of the target

word, suggesting that cultural knowledge may be important to consider in the future.

6 Conclusion

We have presented a framework that combines probabilistic inference with neural contrastive learning to generate novel slang word usages. Our results suggest that capturing semantic and contextual flexibility simultaneously helps to improve the automated generation of slang word choices with limited training data. To our knowledge this work constitutes the first formal computational approach to modeling slang generation, and we have shown the promise of the learned semantic space for representing slang senses. Our framework will provide opportunities for future research in the natural language processing of informal language, particularly the automated interpretation of slang.

Acknowledgments

References

Elton Shah Aly and Dustin Terence van der Haar. 2020. Slang-based text sentiment analysis in instagram. In *Fourth International Congress on Information and Communication Technology*, pages 321–329, Singapore. Springer Singapore. **DOI:** https://doi.org/10.1007/978-981-32-9343-4_25

Pierre Baldi and Yves Chauvin. 1993. Neural networks for fingerprint recognition. *Neural Computation*, 5(3):402–418. **DOI:** https://doi.org/10.1162/neco.1993.5.3.402

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160. **DOI:** https://doi.org/10.1111/josl.12080

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D16-1120

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146. **DOI:** https://doi.org/10.1162/tacl_a_00051

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/D15-1075

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1994. Signature verification using a ''siamese'' time delay neural network. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, pages 737–744, Morgan-Kaufmann. **DOI:** https://doi.org/10.1142/9789812797926_0003

Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. 1995. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16:1190–1208. **DOI:** https://doi.org/10.1137/0916069

Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1*

- *Volume 01*, CVPR '05, pages 539–546, Washington, DC, USA. IEEE Computer Society.

C. Paul Cook. 2010. *Exploiting Linguistic Knowledge to Infer Properties of Neologisms*. PhD thesis, University of Toronto, Toronto, Canada.

Verna Dankers, Marek Rei, Martha Lewis, and Ekaterina Shutova. 2019. Modelling the interplay of metaphor and emotion through multitask learning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2218–2229, Hong Kong, China. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/D19-1227`

Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling language models to fill in the blanks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2492–2501, Online. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/2020.acl-main.225`

Connie C. Eble. 1989. The ephemerality of American college slang. In *The Fifteenth Lacus Forum*, 15, pages 457–469.

Connie C. Eble. 2012. *Slang & Sociability: In-group Language Among College Students*. University of North Carolina Press, Chapel Hill, NC.

Katrin Erk. 2016. What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics*, 9(17):1–63. **DOI:** `https://doi.org/10.3765/sp.9.17`

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, 35(1):61–103. **DOI:** `https://doi.org/10.1162/coli.08-010-R1-07-048`

Renato Ferreira Pinto Jr. and Yang Xu. 2021. A computational theory of child overextension. *Cognition*, 206:104472. **DOI:** `https://doi.org/10.1016/j.cognition.2020.104472`

Elena Filatova. 2012. Irony and sarcasm: Corpus generation and analysis using crowdsourcing. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 392–398, Istanbul, Turkey. European Language Resources Association (ELRA).

Ge Gao, Eunsol Choi, Yejin Choi, and Luke Zettlemoyer. 2018. Neural metaphor detection in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 607–613, Brussels, Belgium. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/D18-1060`

David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. 1992. Using collaborative filtering to weave an information tapestry. *Comunications of ACM*, 35:61–70. **DOI:** `https://doi.org/10.1145/138859.138867`

Jonathan Green. 2010. *Greens Dictionary of Slang*. Chambers, London.

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, Berlin, Germany. Association for Computational Linguistics.

Eric Holgate, Isabel Cachola, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. 2018. Why swear? Analyzing and inferring the intentions of vulgar expressions. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4405–4414, Brussels, Belgium. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/D18-1471`

Herman Kamper, Weiran Wang, and Karen Livescu. 2016. Deep convolutional acoustic word embeddings using word-pair side information. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4950–4954. **DOI:** `https://doi.org/10.1109/ICASSP.2016.7472619`

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *Deep Learning Workshop at the International Conference on Machine Learning*, volume 2.

Vivek Kulkarni and William Yang Wang. 2018. Simple models for word formation in slang. 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 1424–1434, New Orleans, Louisiana. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/N18-1129

William Labov. 1972. *Language in the Inner City: Studies in the Black English Vernacular*. University of Pennsylvania Press.

William Labov. 2006. *The Social Stratification of English in New York City*. Cambridge University Press. **DOI:** https://doi.org/10.1017/CBO9780511618208

Sidney Landau. 1984. *Dictionaries: The Art and Craft of Lexicography*, Charles Scribner's Sons, New York, NY.

Maria Lapata and Alex Lascarides. 2003. A probabilistic account of logical metonymy. *Computational Linguistics*, 29(2): 261–315. **DOI:** https://doi.org/10.1162/089120103322145324

Marc T. Law, Nicolas Thome, and Matthieu Cord. 2013. Quadruplet-wise image similarity learning. In *2013 IEEE International Conference on Computer Vision*, pages 249–256. **DOI:** https://doi.org/10.1109/ICCV.2013.38

Adrienne Lehrer. 1985. The influence of semantic fields on semantic change. *Historical Semantics: Historical Word Formation*, 29:283–296. **DOI:** https://doi.org/10.1515/9783110850178.283

Changsheng Liu and Rebecca Hwa. 2018. Heuristically informed unsupervised idiom usage recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1723–1731, Brussels, Belgium. Association for Computational Linguistics.

Rijul Magu and Jiebo Luo. 2018. Determining code words in euphemistic hate speech using word embedding networks. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 93–100, Brussels, Belgium. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W18-5112

Elisa Mattiello. 2005. The pervasiveness of slang in standard and non-standard English. *Mots Palabras Words*, 6:7–41.

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, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden. 2011. Quantitative analysis of culture using millions of digitized books. *Science*, 331:176–182. **DOI:** https://doi.org/10.1126/science.1199644

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 2786–2792. AAAI Press.

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pages 807–814, USA. Omnipress.

Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. 2016. Learning text similarity with Siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157, Berlin, Germany. Association for Computational Linguistics. **DOI:** https://doi.org/10.18653/v1/W16-1617

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Malvina Nissim and Katja Markert. 2003. Syntactic features and word similarity for supervised metonymy resolution. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 56–63, Stroudsburg, PA, USA. Association for Computational Linguistics. **DOI:** `https://doi.org/10.3115/1075096.1075104`

Zhengqi Pei, Zhewei Sun, and Yang Xu. 2019. Slang detection and identification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 881–889, Hong Kong, China. Association for Computational Linguistics.

Martin J. Pickering and Simon Garrod. 2013. Forward models and their implications for production, comprehension, and dialogue. *Behavioral and Brain Sciences*, 36(4): 377–392. **DOI:** `https://doi.org/10.1017/S0140525X12003238`

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Christian Ramiro, Mahesh Srinivasan, Barbara C. Malt, and Yang Xu. 2018. Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, 115:2323–2328. **DOI:** `https://doi.org/10.1073/pnas.1714730115`

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

DOI: `https://doi.org/10.18653/v1/D19-1410`

Ekaterina Shutova, Jakub Kaplan, Simone Teufel, and Anna Korhonen. 2013a. A computational model of logical metonymy. *ACM Transactions on Speech and Language Processing*, 10(3):11:1–11:28. **DOI:** `https://doi.org/10.1145/2483969.2483973`

Ekaterina Shutova, Simone Teufel, and Anna Korhonen. 2013b. Statistical metaphor processing. *Computational Linguistics*, 39(2):301–353. **DOI:** `https://doi.org/10.1162/COLI_a_00124`

Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 4080–4090.

Ian Stewart and Jacob Eisenstein. 2018. Making ''fetch'' happen: The influence of social and linguistic context on nonstandard word growth and decline. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4360–4370, Brussels, Belgium. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/D18-1467`

Zhewei Sun, Richard Zemel, and Yang Xu. 2019. Slang generation as categorization. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pages 2898–2904. Cognitive Science Society.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. *Metaphor: A Computational Perspective*. Morgan & Claypool Publishers. **DOI:** `https://doi.org/10.2200/S00694ED1V01Y201601HLT031`

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. 2016. Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 3637–3645, USA. Curran Associates Inc.

Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 1386–1393, Washington, DC, USA. IEEE Computer Society. **DOI:** `https://doi.org/10.1109/CVPR.2014.180`

Kilian Q. Weinberger and Lawrence K. Saul. 2009. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics. **DOI:** `https://doi.org/10.18653/v1/N18-1101`

Steven Wilson, Walid Magdy, Barbara McGillivray, Kiran Garimella, and Gareth Tyson. 2020. Urban dictionary embeddings for slang NLP applications. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4764–4773, Marseille, France. European Language Resources Association.