Natural Language Processing for Slang

by

Zhewei Sun

A thesis submitted in conformity with the requirements
for the degree of Doctor of Philosophy

Graduate Department of Computer Science
University of Toronto

Natural Language Processing for Slang

Zhewei Sun
Doctor of Philosophy
Graduate Department of Computer Science
University of Toronto
2024

Abstract

Slang is a common type of language that makes creative and highly flexible use of words. A basic problem that language users tackle is how to develop and interpret novel slang terms for communication in a community. This problem is relevant for natural language processing since new slang expressions often emerge in daily conversations and online social media. However, principled computational approaches to modeling slang are lacking, which presents key challenges to the effective natural language processing of slang. In this dissertation, I develop a computational framework that offers new methodologies for the automated generation, interpretation, and translation of English slang word usages, as well as for characterizing the principles in slang variation across language communities.

My dissertation is organized into three main parts. The first part addresses the under-explored problem of slang semantic extension, namely how existing words in the lexicon take on new meanings in informal context. I develop a generative framework that combines contrastive learning with probabilistic models of semantic chaining to capture slang semantic extension. By leveraging dictionary-based resources of slang, I show how the learned semantic representations more accurately predict slang word choices compared to existing approaches that rely more exclusively on corpus data. The second part of my dissertation tackles the inverse problem of slang interpretation by applying these semantic representations to interpret and translate novel slang usages in natural text. I show how this approach provides better accuracy and sample efficiency in both slang interpretation and translation, in comparison to baseline con-

textualized language models. Finally, the third part of my dissertation investigates semantic variation of slang across different language communities focusing on a comparative study of US and UK. I show that models incorporating either communicative need or semantic chaining can predict the regional identity of slang usages.

In summary, my dissertation contributes a principled framework for modeling the lexical semantics and usages of English slang, and it opens up future opportunities for the computational investigation and automated processing of informal language across a diverse set of communities and languages.

To my father for his unwavering support throughout this journey.

To my aunt for raising me and enabling this monumental achievement of my life.

Acknowledgments

I would like to express my most sincere gratitude to everyone who has supported me during my Ph.D. studies. First and foremost, I wish to thank my doctoral supervisor, Prof. Yang Xu, for providing me with this opportunity and for his dedication and patience throughout my Ph.D. I would not have produced this thesis without your invaluable advice. I am fortunate to be able to tell others, with confidence, that I have enjoyed my Ph.D. studies. Thank you for your kind support.

I would also like to thank my committee members Prof. Graeme Hirst and Prof. Richard Zemel for their continued support, from the inception of my thesis topic all the way towards its completion. Your insightful questions and feedback have been instrumental in enriching the quality of my work.

Furthermore, I would like to thank Prof. Raquel Fernández, for being a wonderful external appraiser and Prof. Michael Garton for chairing my final defense in such a professional manner.

I am indebted to all my professors and colleagues in the Computational Linguistics group who have contributed to improving the quality of my work. Special thanks to everyone in the Language, Cognition, and Computation group for their encouragement and feedback throughout the development of this work. I would also like to thank Jai Aggarwal and Katie Warburton for organizing the student social events. They have kept me mentally nourished during the uncertain days of the pandemic.

I would like to acknowledge OSAP, NSERC, and Amazon Alexa for contributing to the funding of my Ph.D. research via scholarships and research grants.

Last but certainly not the least, I would like to express my deepest gratitude to my family who has made this journey encouraging and comfortable. Thank you all for always believing in me.

Toronto,

November 2023

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 What is slang?

Slang is a type of language that is often used in colloquial speech. The recent emergence of online social media platforms has also made slang more commonplace in written forms. A precise definition of slang, however, remains an open question among linguists (Dumas and Lighter, 1978). For example, the Oxford English Dictionary (OED) defines slang as "language of a highly colloquial type, considered as below the level of standard educated speech, and consisting either of new words or of current words employed in some special sense" (Stevenson, 2010).[1] Meanwhile, Merriam-Webster (2004) defines slang as "language peculiar to a particular group" and that it is "an informal nonstandard vocabulary composed typically of coinages, arbitrarily changed words, and extravagant, forced, or facetious figures of speech".[2] Finally, in her seminal work on slang, Eble (2012) defines slang as "an ever changing set of colloquial words and phrases that speakers use to establish or reinforce social identity or cohesiveness within a group or with a trend or fashion in society at large". Indeed, even the most recent slang definitions from well-known sources vary substantially from each other. As Dumas and Lighter (1978) have pointed out,

[1] https://www.oed.com/view/Entry/181318
[2] https://www.merriam-webster.com/dictionary/slang

such definitions also tend to be imprecise. For instance, what constitutes as "highly colloquial language" and "arbitrarily changed words" is very loosely defined.

Fortunately, many important characteristics of slang are commonly agreed upon. First, slang is often considered informal (Spears, 1981; Landau, 1984; Mattiello, 2005), meaning that it is seldom seen in formal text (Michel et al., 2011). Next, slang is often short-lived (Sornig, 1981; Eble, 1989; Tagliamonte and Denis, 2008) with a tendency to phase out quickly compared to standard language. Also, slang usages are innovative and flexibly employ creative linguistic devices such as metaphor, amelioration, and irony (Warren, 1992; Eble, 2012). Finally, the creation and use of slang are socially driven (Labov, 1972, 2006; Slotta, 2016). For example, community-specific slang can be created and used to reinforce membership status and cohesion within the community (Mattiello, 2005; Eble, 2012). Dumas and Lighter (1978) argue that slang can be defined via a set of core characteristics. Specifically, an expression that embodies many of slang's characteristics is likely to be treated as slang by language users.

Alternatively, linguists have attempted to describe slang by differentiating it from other types of non-standard language with overlapping characteristics, including jargon, dialect, profanity, and cant. For example, Mattiello (2005) differentiates slang and jargon by the level of prestige. Whereas jargon is often used by professionals in formal settings to convey prestige and pretentiousness, slang terms tend to be more familiar and spontaneous. Eble (2012) acknowledges that both slang and jargon can be used by specific groups of users, but unlike jargon, slang usually conveys "feelings, attitudes, and unity of spirit". Such a description is consistent with the characteristics of slang: Although both slang and jargon are socially driven, slang tends to be more short-living, informal, and innovative in its expression.

Another important aspect of slang that complicates its definition is conventionalization — when a slang expression loses its defining characteristics and enters the standard language. For example, using the word *cool* to express 'something good'

was considered as a slang usage a hundred years ago. Although it is still arguably less formal than the word *good* today, the use of *cool* to express 'something good' is generally considered standard English and not a slang. It is thus also important to consider the temporal aspect of slang when defining it, for that the categorization of language expressions as slang may be restricted to a specific time-frame. Given the ephemeral nature of slang, most slang usages phase out before reaching conventionalization (Eble, 1989). As a consequence, those usages remain to be perceived as slang, although rarely used.

In this dissertation, I use lexicographic resources of slang to define and distinguish between slang and standard language. Specifically, I consider a word-sense[3] pair to be slang if it can be found in a slang dictionary (e.g., Green, 2010). Word senses found in standard dictionaries such as the OED are considered as *conventional senses* if they are not labeled as slang or informal. Note here that neither the word form nor its meaning alone is sufficient in specifying a slang usage. For example, consider the sentence "Good purchase, that jacket is *blazing*". Here, the slang word *blazing* is used to refer to 'First-rate, excellent' instead of its conventional sense 'Burning brightly' (Green, 2010). In this case, the word *blazing* itself cannot distinguish whether it is being used as a slang expression or not. Only when it is used in the slang context involving a purchased jacket, the meaning of *blazing* inferred from the usage context manifests its use as slang. Similarly, the sense 'First-rate, excellent' can be expressed conventionally using words such as *amazing* and *fabulous* but is only considered a slang sense when attached to the word *blazing*.

Therefore, a well-specified slang usage not only requires a word-form (which may not necessarily be a neologism) but also the attached slang sense. Here, the association between the word-form and the slang sense is non-arbitrary. From a communicative perspective, the slang sense needs to be semantically associated with the word that it is attached to, so that listeners can efficiently infer the intended meaning based on

[3] I will use the terms *meaning* and *sense* interchangeably.

their knowledge of the word (Sornig, 1981; Warren, 1992; Eble, 2012).

1.2 Creation of slang

Human speakers create novel slang usages to address communicative need (Sornig, 1981) and/or to reinforce group membership (Eble, 2012). For example, the use of *beast* to refer to 'Subway #2 of NYC' illustrates communicative need for people living in New York City. Meanwhile, the same word can be used to express excellence in the US but is commonly used to refer to criminals in the UK (Green, 2010). As a result, one who uses *beast* to express excellence reinforces group membership by presenting themselves to others as members of the US community. In both cases, the intended meaning must carry appropriate associations with the lexical choice to facilitate efficient communication between speakers and listeners. Given a to-be-expressed meaning such as 'First-rate, excellent', the speaker can employ two distinct strategies in creating a new slang usage. Slang can thus be broadly categorized into two types based on its method of creation. First, a novel slang expression can be coined to express a not-necessarily-new meaning. Also, an existing expression can be taken from the lexicon with a new slang sense attached to the expression. I refer to these strategies as *coinage* and *reuse*. In this section, I describe both generative processes of slang and potential modeling strategies.

1.2.1 Coinage: Word formation

The *coinage* process can be as simple as creating a new acronym, or invoking more intricate word formation processes such as lexical blending (Eble, 2012). For example, the slang *mocktail* is a result of blending constituent words *mock* and *cocktail* with a meaning of 'A non-alcoholic drink' (Green, 2010). Here, the lexical choice *mocktail* has clear morphological associations with its intended meaning. An alternative word with no clear association, such as *modams*, would be less likely to be attached with

the intended meaning, and even less so for arbitrary forms such as *mktla* that is not composed of standard English morphemes. The creation of slang forms can also take phonemic associations into account (Sornig, 1981). For instance, the slang *knees* is an Australian rhyming slang meaning 'Please'.[4] Here, not only that the morpheme *knee* is metaphorically associated to the intended meaning, but the overall similarly in sound also plays a notable role.

Slang forms are also not restricted to single-word expressions but can appear as multi-word phrases. For example, *bird course* is a well-known Canadian slang phrase that refers to 'An easy course'. Here, compositional semantics dictates the intended meaning instead of morphophonemic units of a word. It is conceivable that the metonymy between a bird's flight and easiness for the word *bird*, combined with the literal meaning of *course*, construe the intended slang meaning 'An easy course'.

In both cases, the meaning of the slang expression can be traced back to its constituent units (i.e. phonemes, morphemes, or words). However, such associations also appear in standard language. It is unclear whether novel coinages that are considered slang differ in their underlying word-sense associations. For instance, the conventional blend *brunch* also owes its meanings to its constituents *breakfast* and *lunch*, but whether its meaning is deduced in the same way as slang such as *mocktail* remains an open question.

1.2.2 Reuse: Semantic extension

Aside from coinage, the *reuse* of an existing word or phrase also makes up a significant portion of slang (Warren, 1992; Green, 2010; Eble, 2012). Here, an existing expression is taken from the lexicon with a slang sense attached (e.g., attaching the sense 'First-rate, excellent' to the word *blazing*), where the slang sense differs from existing conventional senses of the word. The phenomenon of slang *reuse* can be naturally viewed as sense extension, a process in which new senses are attached to an

[4] https://en.wiktionary.org/wiki/Appendix:Australian_English_rhyming_slang

existing word. Paul (1880) describes semantic extension as the derivation of meaning from a word's conventional senses.[5] Here, conventional sense refers to the meaning of a word that is part of the standard language. Language users then derive occasional senses to lexical items. If such derivation receives wide-spread use, then the occasional sense becomes a new conventional sense attached to the lexical item, thus making the lexical item polysemous. For example, *blazing* originally means 'Burning brightly' but has been extended to 'Of outstanding heat' (Merriam-Webster, 2004; Stevenson, 2010). Similarly in the case of slang, the new slang sense can be viewed as an extension from a word's conventional senses. For example, *blazing* extended from 'Burning brightly' and 'Of outstanding heat' to express 'First-rate, excellent' as slang.

To facilitate efficient cognitive processing, words are more likely to extend to senses that are semantically consistent with existing senses (Klein and Murphy, 2001). In the *blazing* example, however, the disparity in meaning appears to be more distant for slang extension. Specifically, there is a smaller semantic gap between the two conventional senses describing fire than between those and the slang sense. Therefore, a natural question to ask here is whether the slang senses emerge from historical conventional senses of the word akin to conventional word sense extension. And if so, in what ways are the mechanisms behind conventional sense extension and slang sense extension similar or different?

A set of 500 conventional and slang usages collected by Warren (1992) shows that the extension devices employed by conventional sense extension indeed differ considerably compared to those of slang. Specifically, the majority of conventional sense extension cases involve particularization of the original sense which makes the resulting sense in close semantic proximity to the original. In comparison, slang semantic extension is much more creative and devices such as metaphor, metonymy, amelioration, and irony are often observed and can appear in conjunction (Eble,

[5] Based on Warren's (1992) account.

Word	Original conventional sense	Extended slang sense	Type of extension
sick	Being ill.	Good, excellent.	Irony
wicked	Evil or mischievous by nature.	Excellent, wonderful.	Irony
slut	A sexually promiscuous woman.	An affectionate term of address among women.	Amelioration
future	An expectation of advancement or progressive development.	An unattractive man.	Pejoration
all-nighter	Something that lasts throughout the whole night.	working all night before an examination.	Particularization
blazing	Burning brightly and with great heat, force, etc.	First-rate, excellent.	Metonomy
kick	Propel with foot.	A strong taste.	Metonymy
beast	A large animal.	A fast car.	Metonymy
wolf	A wild, dog-like animal.	A predatory person.	Metaphor
ice	Frozen water.	To kill.	Metaphor
night owl	An owl (order Strigiformes) that is nocturnal.	Anyone who is habitually out and about at nighttime.	Metaphor
steamed	Being vaporized.	Being angry.	Metaphor

Table 1.1: A list of slang usages from Green (2010) and Eble (2012) illustrating the rich set of innovative sense extension strategies that are employed in slang word reuse.

2012). Table 1.1 shows a list of example slang usages involving such sense extension strategies.

Warren analyzed 1,000 sense extension examples and categorized them into a pre-defined set of sense extension devices. Inspired by prominent theories in the field (Paul, 1880; Stern, 1931; Ullmann, 1942), Warren categorized sense extension into the following four broad categories:

- **Particularization** — A hyponymic sense with additional distinctive features added to the original sense. E.g., *cutting gear*: 'Equipment for cutting' → 'Gas-operated cutting equipment that breaks into safes'.

- **Implication** — A novel sense formed by retrieving additional contextual information from relevant communicative contexts. E.g., *sweat*: 'Emit sweat' → 'Work hard'.

- **Metonymy** — Formed by aggregating features of referents that are *closely connected* to the original sense. E.g., *gate*: 'Structure to block entrance' → 'Money

Type	Conventional	Slang
Particularization	169 (33.8%)	56 (11.2%)
Implication	32 (6.4%)	65 (13.0%)
Metonymy	27 (5.4%)	41 (8.2%)
Metaphor	205 (41.0%)	245 (49.0%)
Other	67 (13.4%)	93 (18.6%)

Table 1.2: Distribution of conventional and slang sense extension types among the 1000 sense extension examples from Warren (1992).

collected at gate'.

- **Metaphor** — Formed by aggregating features of referents that are *reminiscent* of the original sense. E.g., *gate*: 'Structure to block entrance' → 'Switch'.

Table 1.2 shows the distribution of both conventional and slang sense extensions categorized among the four proposed sense extension types. The study confirmed that slang senses, like their conventional counterparts, indeed relate to their respective original senses. Furthermore, the underlying devices of sense extensions are closely shared, where the same four sense extension devices accounted for a similarly large portion of both conventional and slang senses. However, Warren also pointed out differences between conventional and slang sense extension in the manner in which the sense extension devices are made use of. Namely, the frequency distribution of the sense extension devices show substantial differences between the two while devices such as metaphor tend to have a clear tendency of exaggeration when used in slang extension.

By examining a collection of American campus slang collected from 1972 through 1993 from The University of North Carolina (UNC) at Chapel Hill, Eble (2012) also described slang as a series of sense acquisition and argues that "it is not merely random but a cognitively guided phenomenon" where newly acquired senses are semantically associated with established senses to facilitate communication. The implications of these findings are twofold. First, similarity in the set of employed sense extension devices suggests that existing computational models of semantic extension can serve as a good starting point in modeling slang. At the same time, there exist differ-

ences between how such devices are applied and the extent to which two senses are considered as relatable.

1.2.3 Modeling slang

Given the different methods in which a slang usage can be created, one can create principled computational models of slang that reflect its generative process. I now discuss the extent to which existing NLP methods can tackle this challenge and potential limitations with existing methods.

For slang coinage, existing work in NLP has proposed deep learning based models to predict word formations based on its constituents (Kulkarni and Wang, 2018). For example, predicting the word *mocktail* given *mock* and *cocktail*. However, the semantic associations between word forms and slang senses have not been explicitly modeled. Nevertheless, prominent word formation strategies such as lexical blending have received careful treatment in the literature (Cook and Stevenson, 2010b; Deri and Knight, 2015; Gangal et al., 2017; Pinter et al., 2020). Also, existing NLP models can treat slang coinages as out-of-vocabulary words (OOVs) and a myriad of techniques have been proposed to produce semantic representations of OOVs (e.g., Sennrich et al., 2016; Pinter et al., 2017; Cotterell and Schütze, 2018; Kudo, 2018; Kudo and Richardson, 2018). Although these techniques have not been systematically evaluated on slang, they can be readily applied to obtain a semantic representation for a newly coined slang expression. I review this body of work in Section 2.3.

Unlike the case of coinage, no existing work in NLP has explored the word-meaning associations in slang reuse.[6] Furthermore, simply applying existing models of semantic extension for conventional language change will be limited. The key limitation lies in the representation of senses using standard sentence embedding techniques that only capture surface-level similarities. For example, a BERT-based embedder (Devlin et al., 2019; Reimers and Gurevych, 2019) would assign similar embeddings to

[6] In the case of slang reuse, both words and phrases are treated in the same way. That is, both words and phrases are characterized by the conventional senses attached to them.

conventional senses of *blaze* 'Of outstanding heat' and 'Burning brightly' because both senses refer to similar concepts describing flame. However, when those conventional senses are compared to the slang sense 'First-rate, excellent', the metonymy manifests a larger semantic gap between the senses which will cause the embedding model to assign embeddings that are further apart. But since slang has a tendency to extend senses in more innovative ways, the models can be substantially improved by accounting for such behavior. In the case of the *blazing* example, if metonymy is more common in slang sense extension, then sense pairs that reflect such associations should be considered semantically close.

Using data from Warren (1992), I show in Chapter 3 that semantic representations from off-the-shelf NLP representations indeed produce higher semantic distances between sense pairs from slang sense extension compared to those in conventional sense extension. To address this gap in knowledge, this dissertation focuses on the modeling of slang reuse as semantic extension. I show such a semantic model of slang reuse can be applied to core tasks of slang in NLP.

1.3 Slang in natural language processing

Processing of slang remains a difficult challenge for many commercial systems today. For example, using Google Translate on the sentence "It makes me *steamed* when I run out of money" would incorrectly output 'Being vaporized' instead of the correct meaning 'Being angry' for the slang *steamed*. In this case, the system falls short because it fails to recognize the correct meaning of the slang *steamed*. To successfully address a task such as the machine translation of slang, the system needs to interpret the correct meaning of *steamed*. At the same time, it is also desirable for the translation system to express the concept 'Being angry' naturally in the target language, ideally generating a slang equivalent. In both cases, the system should have the ability to detect the slang usage so that processing can be triggered on appropriate spans

of text. Therefore, a successful system must have the ability to perform well in the following core tasks:

1. **Slang Detection:** Given a sentence (e.g., "It makes me steamed when I run out of money"), detect whether a slang is being used in the sentence. If so, identify the exact position of the slang expression (e.g., *steamed*). This allows a downstream task to perform slang-specific processing on the detected expression.

2. **Slang Interpretation:** Given an identified slang within a sentence (e.g., "It makes me *steamed* when I run out of money"), interpret its intended meaning within the usage context (e.g., 'Being angry'). This allows the system to obtain the intended meaning of the slang.

3. **Slang Generation:** Given an intended meaning to be expressed (e.g., 'Being angry'), choose a word or phrase (e.g., *steamed*) to express the intended meaning as slang. In the case of machine translation, this allows the system to generate an equivalent slang usage in the target language.

Also, these tasks are not specific to machine translation but are applicable to a broad spectrum of important NLP tasks. For instance, a sentiment analysis system would require both detection and interpretation to identify the correct sentiment connotation of a slang usage. A dialogue system, on the other hand, needs both detection and interpretation[7] when acting as a listener, and proceeds to generation when it is its turn to speak.

Despite the importance of these tasks, existing work on the natural language processing of slang has been focusing primarily on slang detection (Pal and Saha, 2013; Pei et al., 2019). In the existing slang processing systems, the slang is processed primarily using its usage context. In other words, relying on distributional semantics (Firth, 1957; Deerwester et al., 1990; Erk, 2016) to infer the meaning of the

[7]Note that it is possible for an NLP system to perform some of these tasks simultaneously. For instance, a system can detect slang by simply interpreting the intended meaning of every word/phrase in a sentence. However, it is possible to detect slang without a precise interpretation. For example, by matching Part-of-Speech-shift patterns that are frequently employed in slang but not conventional language (Pei et al., 2019).

slang (Ni and Wang, 2017). Under this paradigm, both generation and interpretation of slang are difficult because distributional semantics alone is insufficient in capturing slang. Consider the sentence "I have a feeling he's gonna [MASK] himself someday". Directly applying a large 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, e.g., *ice*—a slang alternative for *kill* (Green, 2010)—received virtually zero probability, suggesting that existing models of distributional semantics, even the transformer-type models, do not capture slang effectively, if at all. While such discrepancies in likelihood may be an indication of surprisal and thus beneficial in slang detection, it would be difficult to either generate or interpret the slang *ice* without an accurate semantic representation that links *ice* with its intended meaning 'To kill'. For this reason, existing methods rely on the distributional context alone to make an interpretation and forgo information encapsulated in slang terms: While it is possible to infer the meaning from the context sentence "I have a feeling he's gonna [MASK] himself someday", the semantic association between 'To kill' and the slang word *ice* (in this case, a metonymy relation) has been disregarded.

For both slang generation and slang interpretation, a semantic representation that captures the nuanced relations between the conventional and slang meanings of an expression (e.g., 'Frozen water' and 'To kill' respectively for the word *ice*) is thus crucial. Using deep learning based embedding methods, I reinforce the findings of Warren (1992) that slang semantic extension treats sense association differently compared to conventional sense extension. Because of this, NLP models trained on conventional language (including models of distributional semantics) will not generalize well on slang as they only capture surface-level similarities between two meanings. I address the under-representation problem by applying contrastive learning (Baldi and Chauvin, 1993; Bromley et al., 1993; Chopra et al., 2005; Weinberger and Saul, 2009) on a large collection of slang dictionary entries to automatically extract patterns of seman-

tic extension attested in the data. In this dissertation, I show how such a semantic representation of slang can be applied to enhance both generation (Chapter 3) and interpretation (Chapter 4) of slang under a principled computational framework.

The discussion thus far assumes that slang usages are universally defined for all language users. However, the use of slang varies substantially across different groups of users. For example, a significant portion of slang usages are regional (Green, 2010), necessitating modeling approaches that produce results tailored to each individual region. In Chapter 5, I show that slang varies much more across different regions compared to conventional language. However, the cause and behavior of such variation in slang usage are not well understood at a macroscopic level. That is, existing linguistic work often delineates a small but carefully studied sample of slang within individual communities of interest (e.g., Denis, 2021), but how slang behaves as a whole is poorly understood. Recent interest in social media analysis has expanded the scope to explore large samples of slang usage in online settings, studying two modes of variation pertinent to slang:

1. **Lexical Variation:** The difference in *word choice* for the same to-be-expressed meaning. For example, both *blazing* and *massive* are used to express excellence, but *blazing* is typically used in the US and *massive* is used in the UK.

2. **Semantic Variation:** The difference in *meaning* expressed by the same word form. For example, the same word *blazing* means 'First-rate, excellent' in the US but is used to express 'Being angry' in the UK.

Prior work on slang variation has primarily focused on lexical variation of online language (Altmann et al., 2011; Eisenstein et al., 2014; Nguyen et al., 2016; Del Tredici and Fernández, 2018; Stewart and Eisenstein, 2018). Work on semantic variation of slang has been sparse in comparison (Del Tredici and Fernández, 2017; Lucy and Bamman, 2021; Keidar et al., 2022). Here, existing work focuses on the quantification of semantic variation and/or determining its causes but does not describe or model

the underlying processes. Chapter 5 of this dissertation shows how a semantic model of slang can be applied to directly model slang semantic variation. By doing so, I show how the model can be applied to predict the regional identity of slang while also being able to reveal theoretical insights on semantic variation of slang.

1.4 Key modeling challenges

Processing of novel slang is no easy task even for humans, with studies showing that interpretation (Braun and Kitzinger, 2001) and translation (Mattiello, 2009) of unfamiliar slang to be much more difficult than conventional language. Although machines have the advantage of observing much more data than humans, the automatic processing of slang remains a challenging task due to slang's inherent characteristics. Here, I elaborate on how defining characteristics of slang makes its modeling practically challenging in NLP.

First, the use of slang is innovative and involves a rich set of generative strategies including but not limited to metaphor, metonymy, amelioration, and irony (Warren, 1992; Eble, 2012). Each extension strategy by its own right raises challenging problems in natural language processing (e.g., Cook and Stevenson, 2010a; Veale et al., 2016; Magu and Luo, 2018). Also, distributional semantics models such as Word2Vec (Mikolov et al., 2013) and BERT (Devlin et al., 2019) are trained primarily on literal text and thus only capture surface-level similarities between meanings. A pair of metaphorically related senses, for example, may not be considered similar as senses that have distant surface meanings would have been seen in very different contexts during training.

Next, newly created slang usages tend to be ephemeral (Sornig, 1981; Eble, 1989; Tagliamonte and Denis, 2008). As a result of this, the set of slang in which an NLP system needs to handle can change rapidly in a short amount of time. Therefore, a practical system has to be capable of processing slang that is potentially unseen

or rarely observed in training data, often resulting in a few-shot learning scenario. Many existing systems that rely on lexicons built on slang dictionaries (e.g., Pal and Saha, 2013; Dhuliawala et al., 2016; Gupta et al., 2019; Wilson et al., 2020) are therefore insufficient in this regard. Similarly for deep learning based methods, data memorization cannot generalize to novel slang. Although large language models such as GPT-3 (Brown et al., 2020) can memorize slang usages seen during training, the handling of novel slang would require the model to be continuously trained and would incur high financial and environmental cost (Bender et al., 2021).

Another characteristic of slang affecting practical systems is its informality (Spears, 1981; Landau, 1984). Existing NLP approaches rely on large training corpora derived from sources such as Wikipedia[8] and Common Crawl[9] which are composed of formal documents such as news articles and wiki pages. Because of this, models observe formal language use much more frequently during training. Informal language such as slang is thus severely under-represented in the data and the result is twofold. First, when facing a tradeoff between performing well on formal versus informal language, a general purpose NLP model will prioritize the former during training because formal language makes up a much larger proportion of the training data. However, finetuning a model directly on informal usage will likely result in much inferior performances on formal language. Second, examples on informal uses of language are scarce even though large neural network based models require large amounts of data to learn effectively (Thompson et al., 2020). This either makes the automatic processing of slang a low-resource problem which necessitates efficient learning or forces the use of large but noisy data sources such as the Urban Dictionary (cf. Swerdfeger, 2012; Nguyen et al., 2018).

Finally, the creation and use of slang are contextually motivated and often reflect group membership (Labov, 1972; Mattiello, 2005; Labov, 2006; Slotta, 2016), a hallmark feature of slang distinguishing it from other types of informal language.

[8] http://en.wikipedia.org
[9] http://commoncrawl.org

This may include, for example, the need to communicate a particular concept in a social group (Sornig, 1981) or the creation of new language to reinforce group cohesion (Mattiello, 2005; Eble, 2012). It is thus important but also potentially beneficial for an NLP system to account for these auxiliary features. For example, if the demographic identity of a user is known to the system, the system may provide a response tailored to the particular identity of the user (e.g., the same slang word *blazing* means excellent in the US but is used to express anger in the UK). Apart from the utility aspect, it is also important to consider potential biases and harm a system may introduce as a result of such distinction. For example, if the NLP system performs well in interpreting *blazing* when it is used to express excellence but not anger, then the system has a performance bias against UK users.

In summary, an effective NLP system for slang addressing these challenges should have the following characteristics:

1. **Representation** of slang semantics: Representing the semantic extension from existing conventional senses to a slang sense, extending the system's capability to allow the processing of both cases of coinage and reuse.

2. **Generalization** to novel slang: The system should be able to process slang that's potentially novel, meaning that it is not already captured or rarely seen in an existing database.

3. **Efficiency** in learning slang: Data sources containing large numbers of high-quality slang usages are often not readily available. An effective NLP approach needs to extract information efficiently in a low-resource setting.

4. **Contextualization** of slang: The NLP system should be aware of the context in which the slang is being used. For example, the same slang used in different socio-demographic context may convey different meanings.

This dissertation will hence describe principled approaches that integrate the above characteristics into natural language processing systems for slang. Chapter 3 builds a

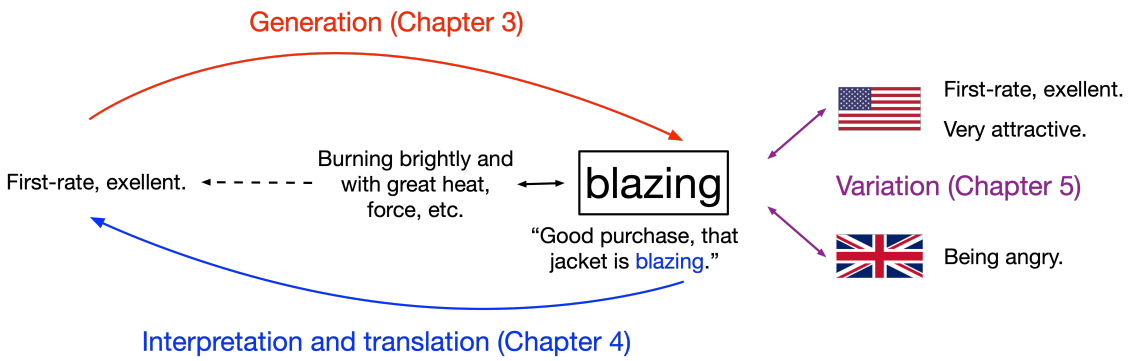


Figure 1.1: The main contributions of this dissertation. Chapter 3 on the generative modeling of semantic extension and word choice of slang; Chapter 4 on applying model of semantic extension of slang for automatic interpretation and Chapter 5 on modeling the regional semantic variation in slang.

computational framework of semantic representation of slang to allow generalization toward unseen slang. Chapter 4 applies the framework to interpret novel slang usages in a data-efficient method. Finally, Chapter 5 considers the contextualization of slang by exploring regional variation in slang semantics.

1.5 Main contributions

This dissertation contributes a novel model of slang semantic extension and shows how such a computational framework can be applied towards the generation, interpretation, and translation of slang and the modeling of semantic variation of slang across different communities. While existing NLP work in modeling word formations can potentially address the processing of slang coinage, no formal models have been proposed to model the semantics of slang reuse. Instead, previous NLP approaches to slang implicitly assume that slang semantics is arbitrary in nature and does not account for the many generative patterns discerned in previous linguistic literature on slang. By adapting models of word sense extension, I show how slang reuse can be computationally modeled as an extension process between a word's original conventional senses and its novel slang senses. Furthermore, the results show that there indeed exist regularities in slang semantic extension that can be learned from data.

Building upon these findings, this dissertation shows how the modeling of slang semantics can benefit important NLP tasks such as the generation, interpretation, and translation of slang which have been sparsely studied due to their sheer difficulty. Finally, I illustrate how the semantic variation of slang can also be modeled as a process of semantic extension, an application of which allows the inference of a slang's regional identity. Figure 1.1 illustrates an example of the processes being modeled.

The modeling of slang as semantic extension improves existing approaches to automatic slang processing which rely heavily on distributional semantics, where the semantic content of the slang expression itself is often ignored. For example, inferring the meaning of the slang *blazing* using only the context sentence "Good purchase, that jacket is *blazing*". Such distributional semantics based models do not fully leverage information encapsulated in the word *blazing* where its slang sense shares semantic association with conventional senses of *blazing*. In the simplest application of distributional semantics, the model makes no distinction between the word *blazing* and any other alternatives, even if the alternative slang words convey very different meanings. My work shows how a standard distributional semantics based embedding model can be warped to capture patterns of slang sense extension by applying efficient contrastive finetuning on dictionary data. I show how the learned embeddings capture semantic relations beyond surface-level similarities reflected in existing distributed semantic models and how such representations lead to substantial improvements in both automatic generation and interpretation of novel slang.

1.6 Overview of dissertation

Chapter 2 reviews the existing literature on work related to NLP for slang to provide a bird's-eye view of the existing capabilities and shortcomings. I will also review research areas that are closely related to slang and discuss how existing approaches in these areas can be applied to slang.

Chapter 3 addresses the generalization and under-representation of slang by automatically extracting patterns of semantic extension pertinent to slang from a large collection of slang dictionary entries. The resulting semantic representations allow the comparison of senses beyond surface-level similarities. For example, the senses 'Burning brightly' and 'First rate, excellent' are dissimilar in their literal meanings but are brought closer together in the resulting representation space if the association between force and excellence is a commonly observed semantic extension pattern during training. The extracted semantic extension patterns also generalize beyond the usage cases observed in our dictionaries, allowing them to be applied against novel slang usages unseen during training.

Chapter 4 shows a data-efficient approach to zero-shot slang interpretation. I show that by combining both semantic information from a generative semantic model of slang and contextual information from large language models (LLM), it is possible to achieve better predictive accuracy, without task-specific training, than a sequence-to-sequence network (Sutskever et al., 2014) trained on a large collection of Urban Dictionary entries (Ni and Wang, 2017). Furthermore, the inclusion of a semantic model of slang allows better representation and generalization of slang usage and results in improved interpretation accuracy when combined with both the LLM and sequence-to-sequence based contextual interpreters. I also show potential improvements the slang interpreter can make in machine translation of slang.

Chapter 5 explores the contextualization of slang and how it may affect the performance of natural language processing systems. I present a first study in this direction by quantifying and modelling the regional semantic variation that exists in slang usage. The experiments show that both lexical and semantic variation are prominent in slang usage. Furthermore, semantic variation in slang can be explained by functions of communicative need (Sornig, 1981) and semantic distinction (Mattiello, 2005; Eble, 2012) within social groups. I show how such knowledge can allow the automatic inference of a user's demographic region using models of semantic variation for slang.

Chapter 6 gives a summary of the contributions on slang NLP and how they address the outlined challenges. I will also describe exciting and promising avenues of future research in slang NLP.

Chapter 2

Related work

2.1 Overview

In this chapter, I survey the existing literature to bring a birds-eye view of existing NLP work related to slang. Work on natural language processing for slang has been relatively sparse and remains an underrepresented area in the literature. Existing NLP work on slang focuses primarily on its automatic detection without principled modeling of slang semantics. Specifically, these approaches often treat slang as a black box and do not take advantage of the theoretical insights uncovered by previous work. In Section 2.2.1, I review the existing NLP approaches to slang based on both dictionary retrieval and deep learning.

Although slang itself has been relatively understudied in NLP, work in other areas of NLP indirectly relates to or addresses some challenges faced in NLP for slang. For example, existing work in computational social science often studies the variation in online language found in social media. Much of the language studied overlaps with slang and the field has made considerable progress in discerning the linguistic and social variables that dictate the behavior of such language. Section 2.2.2 will review this line of work.

The NLP community has also spent a tremendous effort on accurately representing out-of-vocabulary words (OOVs). This is especially relevant to slang as a good portion

of slang involves coinage (Sornig, 1981). Although this dissertation focuses on the understudied case of reuse, a review of the existing NLP work that is potentially applicable to slang coinage will be given for completeness. Section 2.3 will review existing NLP work on lexical blending, one of the most prominent word formation processes used in slang (Mattiello, 2005; Eble, 2012; Kulkarni and Wang, 2017), as well as general purpose OOV models that do not make specific assumptions about word structure.

Since this dissertation makes extensive modeling of slang-reuse as a phenomenon of semantic extension, I will introduce in Section 2.4 the necessary background and related work on the computational modeling of semantic extension. Here, I focus on semantic chaining models based on cognitive models of categorization, an important methodology that will be applied in later chapters of this dissertation.

Finally, Section 2.5 outlines the available large-scale data resources for slang. In doing so, I discuss the advantages and limitations of such data sources and why certain datasets are selected for both training and evaluations.

2.2 Computational studies of slang

2.2.1 Automatic processing of slang

2.2.1.1 Dictionary-based approaches

Most existing approaches in the natural language processing for slang focus on efficient construction, extension, and retrieval from dictionary-based resources. The resulting slang database is then relied upon for all downstream tasks (e.g., detection, interpretation). However, many entries from large dictionaries such as the Urban Dictionary (UD) may be of poor quality (Swerdfeger, 2012). Therefore, a common theme in such approaches is to control or improve the quality of the database while attempting to generalize its existing entries to the furthest extent.

Pal and Saha (2013) described a dictionary-based slang retrieval system focusing on the task of slang detection. First, they constructed a database of slang the user is interested in detecting. During test time, each incoming word in the text is then checked against all entries in the database, with an additional module that matches sounds-alike words (e.g., *alpa* would be matched against *alpha*). The system assumes that any sounds-alike word being used in a diverse set of contexts is likely a slang. Specifically, if such a word has been used along many different context words, the system adds the novel slang into the slang database as a way to account for unseen slang.

Dhuliawala et al. (2016) proposed an annotation scheme to control the quality of Urban Dictionary entries. In their proposed 'SlangNet' database, only UD entries attested in a Reddit message would be extracted. To do so, sentences are first extracted from a Reddit crawl. For each word in the sentence that does not appear in WordNet (Miller, 1994) (i.e. coinage), the top definition from UD is taken as its definition. For all other content words (i.e. possible reuse), tag words[1] from the corresponding UD entries are matched with conventional senses of the word in WordNet using bag-of-words overlap. If no match has been found, then the same matching process is applied to sense entries of the word in UD. The set of matching UD senses is then used as annotation candidates where the annotators could choose one of the candidate senses or write their own definition.

Wu et al. (2018) constructed a sentiment lexicon of slang by labeling lexical entries found in Urban Dictionary. They begin the sentiment labeling process by beginning with a small set of seed words that are also available in exisiting sentiment lexicons such as SentiWordNet (Baccianella et al., 2010), LIWC (Pennebaker et al., 2001), and MPQA (Deng and Wiebe, 2015). The sentiment scores are then propagated successively using Twitter sentences and UD metadata. Words that co-occur in tweets are assumed to share similar meaning, thus similar sentiment. Similarly, words that

[1] the authors claim that this is a set of words that relates to either the conventional or slang senses of the word.

are labeled as "related" in UD are given similar sentiment ratings. The propagation method also allows the automatic labeling of new word lists from UD, enabling the system to generalize towards unseen words.

Gupta et al. (2019) introduced a fuzzy-logic (Zadeh, 1965) based algorithm to refine definition rankings given in Urban Dictionary. Whereas UD ranks all definitions for a given word solely on the differences between the number of upvotes and downvotes, the proposed SLANGZY algorithm considers additional features such as the definition length and average upvotes across all definitions in a word. The final ranking is then produced by taking a weighted average of the feature scores where the authors recommended a set of manually tuned weights.

The key disadvantage of dictionary-based systems is that they fail to generalize beyond what was available in the training data, even if all dictionary entries are of high-quality. This makes such systems futile in the face of unseen slang. However, one of the key challenges in processing slang is how ephemeral slang usages are (Eble, 1989), making it important for NLP systems to be able to process unseen slang that will not necessarily be accurately recorded in a dictionary in a timely manner. The methods proposed in these dictionary-based approaches are often based on heuristics and have not been evaluated carefully in large-scale experiments.

2.2.1.2 Deep learning based approaches

With the popularization of deep learning, recent NLP work on slang has been tackling the generalization issue using data-driven approaches. Namely, the development of systems that can process unseen slang words or meanings by generalizing examples of slang usage from the training data.

Ni and Wang (2017) formulated English slang interpretation as a translation task (although they did not tackle slang machine translation *per se*). In this work, each slang query sentence in English is paired with the ground-truth slang definition (also in English), and such pairs are fed into a translation model. In addition, the spellings

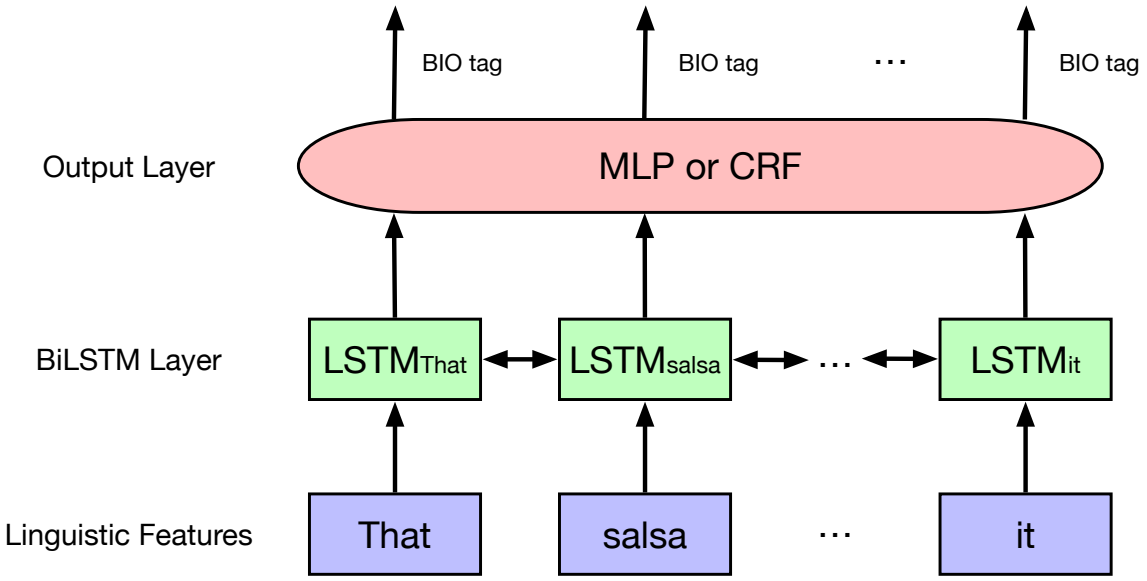


Figure 2.1: Illustration of Pei et al. (2019)'s neural architecture for slang detection.

of slang word forms are also considered as input. The proposed neural network encodes both the usage context and the slang form using separate Long short-term memory (LSTM) encoders (Hochreiter and Schmidhuber, 1997). The two encoded representations are then linearly combined to form the encoded input for a sequence-to-sequence network (Sutskever et al., 2014). During training, the combined state is passed onto an LSTM decoder to train against the corresponding definition sentence. During test time, beam search (Graves, 2012) is applied to decode a set of candidate definition sentences. Yi et al. (2019) showed that a similar architecture can be applied to process Chinese slang. Instead of encoding the slang form's individual characters, phonetics of the words are encoded instead.

Pei et al. (2019) developed a deep learning based approach for automatic detection and identification of slang. Illustrated in Figure 2.1, the proposed end-to-end neural architecture first encodes word-level information and relevant linguistic features using an LSTM network. A conditional random field (Lafferty et al., 2001) is then imposed onto the LSTM network to make sequence-level predictions. They found that while an end-to-end neural-network based approach is effective, linguistic features such as contextual surprisal also play an important role in improving slang detection accuracy.

Wilson et al. (2020) trained fastText (Bojanowski et al., 2017) embeddings by treating Urban Dictionary entries as text documents. The resulting embeddings were shown to capture non-standard semantic relationships between words. In particular, the authors have identified examples in which the model implicitly generalizes towards unseen slang by combining fastText embeddings of the unseen word's constituent subwords. In the case of word reuse, however, the resulting embeddings do not distinguish between conventional and slang meaning and can potentially capture either.

All of these approaches directly apply machine learning models without building an explicit semantic representation for slang. Instead, the models rely on the distributional context. Specifically, Ni and Wang (2017) encoded the surrounding linguistic context directly using an LSTM encoder; Pei et al. (2019) collected information from the context via a conditional random field; And Wilson et al. (2020) directly applied the skipgram objective. However, distributional semantics is insufficient in capturing slang meanings faithfully. In Chapter 3, I show how linguistically and cognitively principled knowledge can be leveraged to explicitly model slang semantics beyond distributional semantics.

Although existing deep learning based methods have shown promise in generalization towards unseen slang, none of these approaches have incorporated architectural changes that are tailored towards slang. The lack of inductive bias makes these approaches data-hungry in a learning setting where high-quality data is scarce. In Chapter 3, I show how semantic knowledge can be efficiently extracted from high-quality dictionary data. Furthermore, I show in Chapter 4 how such knowledge can be applied to efficiently improve the performance of automatic slang interpretation and translation.

2.2.2 Slang variation

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., 2014b), and socioeconomic status (Labov, 1972, 2006) have been shown to play important roles in shaping slang. Recent surge in the interest of online social media text analysis has seen much prominent work in computational social science studying the characteristics of online language use. Online language is especially relevant to slang as language innovations originated from online communities (e.g. subreddits) have been generally regarded as slang (Del Tredici and Fernández, 2018).

Much of the earlier work focuses on the *lexical variation* of online language studying the differences in word choice among different online communities (Altmann et al., 2011; Eisenstein et al., 2014; Nguyen et al., 2016). More recently, it has also been shown that social variables can predict the popularity and dissemination of lexical innovations in online language. Del Tredici and Fernández (2018) studied a collection of 7,962 Internet slang found across 20 Reddit communities. They showed that apart from linguistic features, slang used by innovators who have stronger ties within a community is more likely to be widespread in the community compared to those used by weakly tied users. Stewart and Eisenstein (2018) also examined how both linguistic and social properties of Reddit slang affect their success. Words that can be used in more diverse contexts are found to be more successful in terms of usage frequency. Meanwhile, social ties corresponding to users of the slang are found to be relevant but less significant.

Aside from lexical variation, prior work has also explored the *semantic variation* of online language: Whether different communities adopt different meanings for the same word? Bamman et al. (2014a) proposed a word2vec (Mikolov et al., 2013) based distributed semantic model to automatically capture community-specific meanings of words. The embedding of a word is the result of combining a global representation and a community-specific representation. The latter is trained using only sentences

from the corresponding community while the former uses all available data. The resulting community-specific embeddings allow the differentiation of meaning across different groups of users.

Applying such methodology, studies have been proposed to quantify the amount of semantic variation in online communities. Del Tredici and Fernández (2017) adapted Bamman et al.'s (2014a) distributive embedding model to train community-specific word embeddings for a small set of Reddit communities and quantified semantic variation by comparing cosine similarities between community-specific embeddings for the same word. Lucy and Bamman (2021) extended the previous study to quantify semantic variation of online language in 474 Reddit communities. They compared PMI-based sense specificity of clustered BERT (Devlin et al., 2019) embeddings generated using different contextual instances of a word's usage, along with an alternative strategy that uses BERT to predict word substitutions from the same usage instances (Amrami and Goldberg, 2019). Lucy and Bamman also proposed a regression-based model of semantic variation with community-based features (e.g., community size, network density) as well as topical features derived from Reddit's subreddit hierarchy.

While community-based features are found to be informative in predicting the strength of semantic variation, the above studies do not explicitly model how slang senses vary. Keidar et al. (2022) performed a causal analysis of semantic change of slang using tweets from 2010 to 2020. Slang's usage frequencies were found to change more drastically than those of conventional language while the semantic change for stable senses progresses much slower. In Chapter 5, I detail my work on modeling the driving forces behind semantic variation of slang. Instead of predicting the causes of semantic variation, my work takes a more direct approach by modeling how slang senses vary. I do so by performing an extended analysis studying attested slang usages over the past two centuries instead of focusing on contemporary Internet slang.

2.3 Models of word formation

A substantial effort exists in the NLP community to address the out-of-vocabulary words (OOVs) problem: The processing of words that are not part of a standard vocabulary. Although most existing work does not focus on slang specifically, many methods are potentially applicable to process slang coinage.

2.3.1 Lexical blending

Earlier work in addressing OOVs emphasizes specific word formation processes that are commonly observed. One example of which is lexical blending, a process which, in its simplest form, combines the prefix of a source word (e.g., *br*eakfast) and the suffix of another (e.g., l*unch*) to create a new word form (e.g., brunch). Blends are also of particular interest in the context of slang because it is one of the most frequent word formation processes observed in slang coinage (Mattiello, 2005; Eble, 2012; Kulkarni and Wang, 2017). Work on lexical blending focuses on decomposing the blended word into its source words, thus allowing some degree of explanation in word interpretation.

Cook and Stevenson (2010b) proposed the first statistical approach to automatically identify source words of a lexical blend in English. Given a word w, they first identified all pairs of words w_1, w_2 such that it would be possible to orthographically combine w_1 and w_2 to arrive at w. Each pair of candidate words are then scored using a set of linguistic features including the frequency, length, phonology, and semantic relatedness of the candidate words. Furthermore, the authors applied this approach to automatically detect lexical blends in text. Here, the premise is that feature scores assigned to a blend's source word pair would be much higher than that of a non-blend. Ek (2018) applied a similar approach on Swedish blends using static word embeddings instead as semantic features and found similar sets of features to be effective.

Another line of work applies data-driven approaches to explicitly model the generative process behind blending. The generative model can then be applied to infer the

source words of a blend. Deri and Knight (2015) proposed a finite-state transducer (FST) to model the blending process. Trained using examples from Wikipedia and Wikitionary, the FST predicts 1,000 candidate source pairs with top probabilities for each blend. The resulting pairs are then reranked using logitistic regression with features similar to those of Cook and Stevenson (2010b). Gangal et al. (2017) applied deep learning methods to model the blending process using a sequence-to-sequence with attention based architecture that encodes two source words and decodes the blend. In addition, a noisy-channel model was imposed to incorporate a character-level language model into the framework. Kulkarni and Wang (2018) proposed a simpler neural architecture by modeling blending as a sequence labeling task. In their proposed task, the source words are placed in a string (e.g., "breakfast#lunch") and a simple LSTM network is trained to output a string indicating which characters to keep (e.g., *CCDDDDDDDDDCCCC* for *brunch*, where *C* refers to kept characters and *D* refers to deleted ones). This approach does not perform as well as that of Gangal et al. (2017) but is much more efficient. Aside from blends, Kulkarni and Wang (2018) also proposed neural architectures to model clippings and reduplicatives, both of which are also common patterns of slang form extension (Eble, 2012) but are sparsely studied in NLP.

Pinter et al. (2020) applied BERT (Devlin et al., 2019) based models to automatically segment and recover the source words of blends. They observed that while conceptually similar, BERT represents compounds and blends very differently. Specifically, the semantic similarities between the source and composed words' BERT representations are much smaller for blends, suggesting more complex semantic shift processes. Furthermore, common tokenization schemes used in BERT-based systems (e.g., WordPiece; Schuster and Nakajima, 2012; Wu et al., 2016, BPE; Sennrich et al., 2016) often cannot find the correct segmentation boundary in blends. Even with the correct segmentation, Pinter et al. (2020) showed that automatically recovering source words for a blend remains a very challenging task even with large contextualized mod-

els like BERT.

2.3.2 Modeling out-of-vocabulary words

As an alternative approach to modeling specific word formation processes, one can assume that all out-of-vocabulary words follow some form of compositional structure. Although the learning process becomes inevitably more difficult, models under this paradigm can generalize towards more words instead of confining to certain types of OOVs (e.g., blends).

A common technique is to decompose an OOV word into a set of subwords. Embeddings are then learned for each subword unit and a representation is obtained for an OOV word by additively combining embeddings of all constituent subwords. Botha and Blunsom (2014) decomposed words into morphemes using the automated system Morfessor (Creutz and Lagus, 2007). Similarly, Wieting et al. (2016) constructed OOV embeddings by combining embeddings of all character n-grams found in the word. Sennrich et al. (2016) adapted Byte-Pair Encoding (BPE; Gage, 1994) to automatically construct a vocabulary of subword units in which embeddings will be learned. Instead of determining subword units linguistically, the vocabulary is obtained by finding the most-frequent character sequences in a corpus. This results in a vocabulary of variable length character sequences and each OOV word can be morphologically decomposed into a series of subword units from the vocabulary. Follow-up work (Kudo, 2018; Kudo and Richardson, 2018) extended this idea by incorporating a unigram language model that outputs a probability distribution over all possible subword segmentations. The resulting embedding then takes consideration of all possible segmentations and this has been shown to be an effective regularizer. Provilkov et al. (2020) achieved similar effect with BPE by modifying it into a stochastic procedure. This can be achieved by randomly dropping out a BPE merging step with a small probability.

Instead of additively combining subword representations, methods have been pro-

posed to learn compositional functions to obtain better embeddings. Ling et al. (2015) proposed a simple embedding method by decomposing all words into characters. A bidirectional LSTM encoder then encodes the sequence of character embeddings to decode a resulting word embedding. Bhatia et al. (2016) applied probabilistic graphical modeling by treating embeddings of an OOV word's morphological decomposition as a prior distribution of the output embedding. Variational inference is then applied to the resulting graphic model to optimize the embedding parameters. Cotterell and Schütze (2018) proposed a probabilistic framework that jointly considers the morphological composition and semantic coherence of the resulting composition. In their composition model, seven different composition functions have been experimented with and vanilla RNNs (Elman, 1990) generally achieved the best performance.

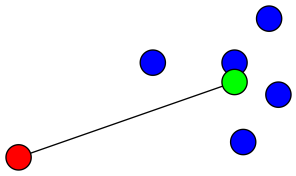
A key limitation of these approaches is that the embedding schemes merely modify the input level of a downstream neural architecture instead of filling in lexical gaps in existing embeddings. This means that all embeddings must be retrained, including those high-quality embeddings from well-known in-vocabulary words. To address this modeling issue, two solutions have been proposed. First, the incorporation of subword information into the word embedding training schemes. The fastText embedding (Bojanowski et al., 2017) does this by injecting subword structure into the skipgram objective. Similar to that of Wieting et al. (2016), each word is decomposed into the set of possible n-gram substrings between length of 3 and 6 inclusive. Vector representation for each n-gram is then checked against those of the context to compute the objective score. An alternative solution proposed by Pinter et al. (2017) is to train a character-level neural network with a set of pretrained embeddings as the network's decoding objective. The network encodes all in-vocabulary words using a bidirectional LSTM encoder and trains the resulting embeddings against a set of pretrained vectors. The character-level encoder would then generalize towards OOV words, allowing embeddings for OOVs to be obtained while retaining the same embedding space from large-scale pretraining.

2.4 Computational studies of semantic extension

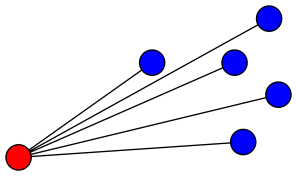
Novel slang senses, like their conventional counterparts, can be viewed as a result of semantic extension from previously attested senses of a given word. Since slang sense extension shares the same mechanisms, existing word sense extension models for conventional language serve as a good starting point in modeling slang reuse.

Earlier work has explored the automatic identification of novel word senses via outlier detection (Erk, 2006), with later methods attuned to identify specific types of change such as widening/narrowing (Sagi et al., 2009) and amelioration/pejoration (Cook and Stevenson, 2010a). Cook and Hirst (2011) constructed synthetic examples to evaluate novel sense identification of infrequent senses and proposed novel similarity and dissimilarity metrics to quantify differences in word senses across two corpora. Follow-up work applied both word-sense disambiguation (WSD) and word-sense induction (WSI) methods to automatically identify novel sense usages (Bamman and Crane, 2011; Lau et al., 2012; Cook et al., 2013, 2014; Mitra et al., 2014, 2015). For example, Lau et al. (2012) applied an unsupervised topic model and Mitra et al. (2014) used graph clustering to cluster the usage instances of a word by the invoked sense.

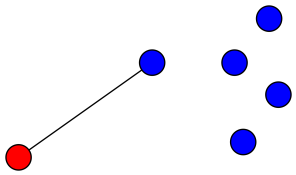
Aside from detection, the extent of semantic change over time has also been quantified. Using Google Ngram (Michel et al., 2011), Gulordava and Baroni (2011) presented the first large-scale quantification of semantic change. In their experiment, a co-occurrence matrix is constructed using bigram context and the resulting entries are used to measure semantic similarity, capturing distributional semantic information about a word. The resulting dichronic vectors are then compared to quantify the extent of semantic change. Kulkarni et al. (2015) combined frequency, syntactical, and word embedding based features to perform change point detection in discerning change in word meaning across time. Later work extended such methodology to discover common patterns in sense extension across history (Xu and Kemp, 2015; Hamilton et al., 2016).



(a) Prototype



(b) Exemplar



(c) One Nearest Neighbor (1NN)

Figure 2.2: Illustration of categorization models. Red (bottom-left) dot denotes novel sense. Blue dots denote existing senses of a candidate word. Green dot denotes prototype (or mean) of the existing senses.

The sense extension process in conventional sense extension has also been explicitly modeled, for example, using graph based methods (Mitra et al., 2014, 2015) and Bayesian modeling (Frermann and Lapata, 2016). More recently, cognitively principled models of semantic extension have been proposed (Ramiro et al., 2018; Habibi et al., 2020). Specifically, a word is considered a linguistic category and its set of senses as members of the category. The association between words and senses is then modeled using cognitively motivated models of categorization.

The premise behind the framework of Ramiro et al. (2018) is that linguistic categories exhibit rich internal structure that resembles human categorization. One such theory is the prototype theory by Rosch (1975), illustrated in Figure 2.2a, in which a prototypical representation is derived from all members in the category. The resulting prototype is then compared to the stimuli to quantify the appropriateness between the category and the stimuli. An alternative theory to that of Rosch is the exemplar

theory by Nosofsky (1986), illustrated in Figure 2.2b. Under the exemplar theory, all members of the category are considered when evaluating a stimuli. Ramiro et al. (2018) also explored the use of One Nearest Neighbor model, illustrated in Figure 2.2c, which resembles Prim's algorithm for finding the shortest distance minimum spanning tree within a graph. It is worth noting that such categorical structure has also been incorporated into recent deep learning architectures for few-shot learning, including the use of both One Nearest Neighbor (Vinyals et al., 2016) and prototype (Snell et al., 2017).

In the context of word sense extension, the stimulus is a novel sense in question and each category represents an existing word form. Within each category, its members represent the existing repertoire of senses the word has acquired in the past. Under this framework, an existing word with senses that are more closely related to a new sense is more likely to be extended to express the new sense. This process is known as *chaining* (Lakoff, 1987; Bybee et al., 1994; Malt et al., 1999; Sloman et al., 2001) and has been shown to effectively model various types of linguistic categories (Xu et al., 2016; Ramiro et al., 2018; Habibi et al., 2020; Ferreira Pinto Jr. and Xu, 2021; Grewal and Xu, 2021; Yu and Xu, 2021). Slang sense extension can also be modeled as a process of chaining using models of categorization. However, it would be naive to assume that the similarities between slang and conventional senses can be compared in the same way as in conventional sense extension. I show in the next chapter how this issue can be alleviated by automatically learning slang-specific patterns of semantic extension from slang dictionary entries.

2.5 Slang data sources

The training and evaluation of natural language processing (NLP) systems require large-scale data sources that contain high-quality descriptions of slang usage. In this section, I explore the existing large-scale digitized slang data sources and discuss

Dataset	# entries	Definitions	Context	Demographics	Releasable
The Online Slang Dictionary (OSD)	11,021	Yes	Yes	No	No
Green's Dictionary of Slang (GDoS)	67,026	Yes	No	Yes	No
Urban Dictionary (UD)					
- Kaggle	2,580,925	Yes	No	No	Yes
- Ni and Wang (2017)	982,281	Yes	Yes	No	Yes
Reddit Glossaries (Lucy and Bamman, 2021)	4,189	Yes	No	Yes	Yes

Table 2.1: Summary of datasets for English slang in natural language processing, including the availability of definition sentences, usage contexts, demographic tags, literal paraphrases, and whether the dataset can be publically released.

their advantages and shortcomings. Table 2.1 summarizes the existing data sources for slang available on the Internet.[2] All the shown data sources are dictionary style datasets where each data entry contains a word/phrase (e.g. *blazing*) and a definition sentence for a slang sense (e.g, 'First-rate, excellent'). In addition, some datasets also provide context sentences in which the slang is being used (e.g. "Good purchase, that jacket is *blazing*") and/or demographic information associated with the slang usage such as year, region, and community of emergence.

Urban Dictionary (UD) is arguably the most well-known data source for slang. UD is advantageous in its sheer size, providing data in a scale that is much larger compared to alternative resources. However, the data is also plagued by poor quality control due to its unmonitored nature. Swerdfeger (2012) outlines potential issues in using Urban Dictionary as an academic source. For example, the only quality control in Urban Dictionary is the user generated upvotes and downvotes. While such voting schemes can work well for well-known slang with widespread public knowledge, it is difficult to achieve consistent quality for less frequent slang used in niche communities since fewer users have the expertise to judge the quality.

Two subsets of the UD have been openly released to the public. The first of which is available on Kaggle.[3] The Kaggle subset is the largest publically available subset of UD containing more than two million entries. Each data entry in this subset also

[2]See Appendix A for more details regarding the access of data.
[3]https://www.kaggle.com/datasets/therohk/urban-dictionary-words-dataset

contains the number of user upvotes and downvotes recorded when the data was retrieved. Such information allows the data entries to be filtered for better quality control. A limitation of the Kaggle subset is that no context sentences are available for the data entries. Ni and Wang (2017) provide a smaller subset of UD[4] containing original context sentences provided by UD users. However, the provided dataset does not contain user votes and UD context sentences tend to be of poor quality.

Alternative online dictionary resources such as The Online Slang Dictionary (OSD) and Green's Dictionary of Slang (GDoS; Green, 2010) contain higher quality data entries but are much smaller in size. For OSD, user submitted entries are manually reviewed by the site administrator and GDoS is an authoritative dictionary authored by a professional lexicographer. Fortunately, these datasets are sufficiently large for training and evaluating models for many important tasks such as the detection (Pei et al., 2019), generation (Chapter 3) and interpretation (Chapter 4) of slang. Although GDoS does not provide context sentences,[5] each slang entry is attached with citations containing year and region of usage. In Chapter 5, I show how such demographic information can be used to analyze variation of slang throughout history.

Recent work by Lucy and Bamman (2021) has also published a glossary of Reddit slang used by different subreddit communities.[6] I do not make use of this resource in this dissertation because the majority of contained entries are acronyms.

Experiments in this dissertation use both OSD and GDoS as the primary data sources to ensure data quality. The regional-historical aspect of GDoS also enables historical analysis of slang and its semantic variation. Chapter 3 also describes a procedure to enhance the quality of UD to be used for training and evaluation. The larger UD sets are only used when the scale is necessary to train neural architectures such as a Sequence-to-Sequence baseline (Sutskever et al., 2014).

It is worth noting that conventional dictionaries such as OED (Stevenson, 2010)

[4] http://www.cs.ucsb.edu/~william/data/slang_ijcnlp.zip

[5] Example context sentences can be automatically inferred from an entry's citations for many definition entries but the sentences are not provided directly by the dictionary.

[6] https://github.com/lucy3/ingroup_lang

also tag certain definition entries as slang. However, I do not use such data because the tagging procedure tends to be inconsistent across dictionaries and even within different editions of the same dictionary (Dumas and Lighter, 1978). Older slang dictionaries such as Flexner (1960) are also not considered because they are either 1) not in digitized form and/or 2) not collected recently. Recall that the definition of slang is a time-sensitive matter due to its ephemerality and conventionalization (Eble, 1989). To avoid these issues, I only consider slang dictionaries that have been updated in the last 10 years.

Chapter 3

Slang generation

The contents of this chapter are based on my previous publication (Sun et al., 2021).

3.1 Motivation

In this chapter, we present a computational method that automatically extracts patterns of semantic extension from slang dictionaries. We show how the extracted patterns can improve semantic representation of slang in NLP systems and how such representations can be applied to build a computational framework that models the generative process of slang word choice, allowing the automatic machine generation of novel slang usage.

Our goal is to extend the capacity of NLP systems toward slang in a principled framework. Given the existing methods that are potentially applicable to cases of coinage (e.g., Cook and Stevenson, 2010b; Pinter et al., 2017), we focus on modeling the generative process of slang reuse. We illustrate the problem of slang word choice in Figure 3.1. Given a to-be-expressed slang sense such as 'To kill', we ask how we can emulate the speaker's choice of slang word(s) in informal context. We are particularly interested in how the speaker reuses existing words from the lexicon and makes innovative use of those words in novel slang context (such as the use of *ice* in Figure 3.1).

Figure 3.1: 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').

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.[1] From a practical perspective, automated slang generation paves the way for automated slang interpretation. Existing psycholinguistic 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

[1] We experiment on English slang but our methodology is applicable to other languages as well.

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 predicts an appropriate word for expressing a query slang meaning given its context, 2) a sense encoder that captures slang meaning in a modified embedding space, and 3) a prior that incorporates different forms of context.

We operationalize our sense encoder using contrastive learning, 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-style 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.

Specifically, the contrastive 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, any conventional and slang sense pairs of *ice*, such as 'Frozen water' and 'To kill', will be encouraged to be in close proximity in the learned embedding space. Moreover, the resulting embedding space will also set apart slang senses from unrelated conventional senses (e.g., pushing away 'To kill' and 'Take a break' where the latter is not a conventional sense of *ice* or its conventional synonyms). As a result, distances between corresponding conventional and slang senses of a word

would be much smaller than those between two unrelated senses. 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. We show sampling strategies that can provide further data augmentation for this data-scarce task. After training, the resulting learned semantic space will be sensitive to common semantic extension patterns of slang.

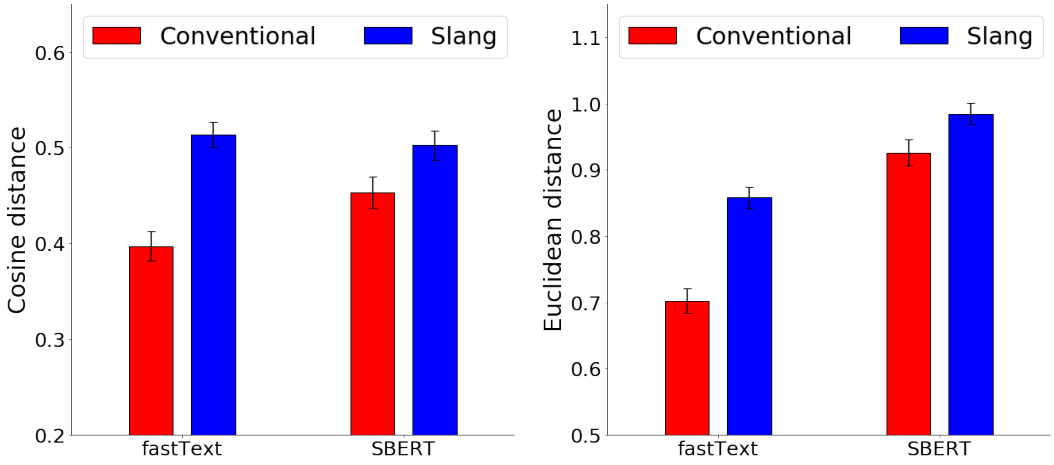
Our framework also captures the flexible nature of slang usages in natural context. Here, we focus on syntax and linguistic context, although our framework should allow for the incorporation of social or extra-linguistic features as well. Recent work has found that the flexibility of slang is reflected 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 3.1). 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. This module allows us to select words that are more likely to be selected as extension candidates with respect to usage context.[2] We also show how a contextualized language infilling model can provide additional prior information from linguistic context (cf. 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.

3.2 Preliminary analysis

Our basic premise is that the word-sense associations between an extended slang sense and the extended word are not random. Specifically, slang word choice depends on

[2]We acknowledge that many slang senses relate to taboo topics such as drug and sex (Green, 2010; Eble, 2012) Since we model the generative process of slang from a speaker's perspective, we assume the to-be-expressed sense is given which makes such biases irrelevant when predicting slang word choice.



(a) Mean cosine distance

(b) Mean euclidean distance

Figure 3.2: Mean sense embedding distances between pairs of conventional and slang sense extensions found in Warren (1992).

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 coldness of one's remains. However, 'Frozen water' and 'To kill' convey very different meanings and we postulate that the innovative nature of slang makes such distant meaning pairs more acceptable than in those of conventional sense extension. In other words, slang sense extensions with innovative characteristics are more likely to be accepted by language users as slang.

Warren's (1992) study has provided preliminary evidence to this claim by contrasting the sense extension strategies employed by conventional and slang sense extension (See Section 1.2.2). To further validate this claim, we measure the semantic distances between pairs of conventional and slang sense extensions found in Warren. This contains 522 instances of slang sense extensions and 488 instances of conventional sense extensions. For each sense, we embed the sense definition sentence using both fastText (Bojanowski et al., 2017) and Sentence-BERT (SBERT; Reimers and Gurevych, 2019). For fastText, we average the word embeddings of all content words within the sentence. For SBERT, we apply it directly on the definition sentence. For each sense extension pair, we measure the semantic distance by computing the cosine and

Euclidean distances between the two sense embeddings.

Figure 3.2 shows the mean embedding distances between pairs of conventional and slang sense extensions. For both fastText and SBERT based definition sentence embeddings, the semantic distances corresponding to slang sense extensions are much greater than those of conventional sense extension. The dichotomy between the semantic distances illustrates the distinction between conventional and slang sense extension and necessitates tailored semantic representations for slang.

A principled semantic representation should adapt to such associations. Our proposed framework is aimed at encoding slang that relates informal and conventional word senses, hence capturing semantic similarities beyond those from existing language models. In particular, BERT-based systems 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.

3.3 Slang generation framework

Our computational framework for slang generation comprises three interrelated components: 1) A probabilistic formulation of word choice 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., 1993)—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.

3.3.1 Probabilistic word choice model

We first introduce a model of word choice adapted from previous work on word sense extension (Ramiro et al., 2018; Habibi et al., 2020), based on the premise that word senses extend by relating new meanings to the current meanings of words that are

closely related in semantic space.

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

$$\begin{aligned} P(w|M_S, C_S) &\propto P(M_S|w, C_S)P(w|C_S) \\ &\propto P(M_S|w)P(w|C_S) \end{aligned} \tag{3.1}$$

Here, we define the prior $P(w|C_S)$ based on regularities of syntax and/or linguistic context in slang usage (described in Section 3.5). We formulate the likelihood $P(M_S|w)$[3] 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 measuring 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:

$$\begin{aligned} P(M_S|w) &:= P(M_S|\mathcal{M}_w) \\ &\propto f(\{sim(E_S, E_{w_i}); E_{w_i} \in \mathcal{E}_w\}) \end{aligned} \tag{3.2}$$

Here, $f(\cdot)$ is a function with range $[0, 1]$ that measures the cohesiveness between a slang sense and a set of conventional senses. 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 3.4, and we compare this proposed method with baseline methods that draw embeddings from existing sentence embedding models.

Our choice of the function $f(\cdot)$ is motivated by prior work on word sense extension (Ramiro et al., 2018; Habibi et al., 2020). Specifically, we consider variants of

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

two established methods in machine learning: 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:

$$f_{1nn}(E_S, \mathcal{E}_w) = \max_{E_{w_i} \in \mathcal{E}_w} sim(E_S, E_{w_i}) \tag{3.3}$$

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:

$$\begin{aligned} f_{prototype}(E_S, \mathcal{E}_w) &= sim(E_S, E_w^{prototype}) \\ E_w^{prototype} &= \frac{1}{|\mathcal{E}_w|} \sum_{E_{w_i} \in \mathcal{E}_w} E_{w_i} \end{aligned} \tag{3.4}$$

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

$$sim(E_S, E_w) = \exp\Big(- \frac{||E_S - E_w||_2^2}{h_s} \Big) \tag{3.5}$$

Here, h_s is a learned kernel width parameter.

A previous iteration of our work (Sun et al., 2019) has directly applied such models along with off-the-shelf word embeddings (Bojanowski et al., 2017) to capture slang word choice. However, a critical limitation is that slang senses, unlike their conventional counterparts, can often be far apart in a vanilla embedding space. As a result, senses that should be considered close together in the context of slang generation (e.g. 'Frozen water' and 'To kill') will not match closely. Section 3.4 will describe an alternative sense embedding scheme that addresses this limitation.

3.3.2 Collaborative filtering

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 (Lehrer, 1985; Xu and Kemp, 2015). For example, the word *snow* may also be a good candidate word to express 'To kill', given its similarity with the true slang *ice* in conventional meanings. We operationalize this by summing over the close neighborhood of candidate word $L(w)$:

$$P(w|M_S, C_S) = \sum_{w' \in L(w)} P(w|w')P(w'|M_S, C_S) \tag{3.6}$$

Here, $P(w'|M_S, C_S)$ is a fixed term calculated identically as in Equation (3.1) 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:

$$P(w|w') \propto sim(w, w') = \exp\Big(-\frac{d(w, w')}{h_{cf}}\Big) \tag{3.7}$$

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

3.4 Contrastive sense encodings

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 *ice* such as 'Frozen water' can hardly be related, in a literal sense, to the slang sense of *ice* such as 'To kill'. 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, since metaphor is one of the frequently employed

sense extension devices in slang, 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., 1993; 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 3.6.4 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.

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

$$E_S = g(M_S); E_P = g(M_P); E_N = g(M_N) \tag{3.8}$$

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:

$$L_{triplet} = \left[m + ||E_S - E_P||_2^2 - ||E_S - E_N||_2^2\right]_+ \tag{3.9}$$

To train the triplet network, we build data triplets from every slang lexical entry (i.e. a word-sense entry in a slang dictionary) in our training set. For each slang sense M_S of word w in a slang dictionary, we create a positive pair with each conventional sense

M_{w_i} of the same word w found in a conventional dictionary. 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.

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.

3.5 Contextual prior

The final component of our framework is the prior $P(w|C_S)$ (see Equation (3.1)) that captures flexible use of slang words with regard to syntax and distributional semantics. For example, slang exhibits flexible Part-of-Speech (POS) shift, e.g., noun→verb transition as in the example *ice*,[4] and surprisals in linguistic context, e.g., *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

[4]Here, *ice* can also be used conventionally as a verb but we consider its most frequent conventional sense for this example. We later show how the frequency distribution across different POS tags can be accounted for.

in the data.

3.5.1 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 (see Section 3.6.2 for details). 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:

$$P(w|C_S) = P(w|P_S) \propto \exp\Big(- KL(\mathcal{P}_w, \mathcal{P}_S^*)\Big)^{\frac{1}{2}} \tag{3.10}$$

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.

3.5.2 Linguistic Context Prior (LCP)

We apply a language model P_{LM} to a given linguistic context K_S to estimate the probability of each candidate word:

$$P(w|C_S) = P(w|K_S) \propto P_{LM}(w|K_S) + \alpha \tag{3.11}$$

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 3.6.3.

3.6 Experimental setup

3.6.1 Lexical resources

We collected lexical entries of slang and conventional words/phrases from three separate online dictionaries: 1) Online Slang Dictionary (OSD),[5] 2) Green's Dictionary of Slang (GDoS) (Green, 2010),[6] and 3) an open-source subset of Urban Dictionary (UD) data from Kaggle.[7] In addition, we gathered dictionary definitions of conventional senses of words from the online version of Oxford Dictionary (OD).[8]

3.6.1.1 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.

[5]OSD: http://onlineslangdictionary.com
[6]GDoS: https://greensdictofslang.com
[7]UD: https://www.kaggle.com/therohk/urban-dictionary-words-dataset
[8]OD: https://en.oxforddictionaries.com

Dataset	# of unique slang word forms	# of slang definition entries	# of conventional definition entries in OD	Avg. definition sentence length
OSD	1,635	2,979	10,091	7.54
GDoS	6,540	29,300	29,640	6.48
UD	1,464	2,631	10,357	9.73

Table 3.1: Summary of dataset statistics for the online slang dictionaries used in the slang generation study.

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 is unique (see Section 3.7.3 for discussion).

3.6.1.2 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.

3.6.1.3 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 GDoS) 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. 5% of the data in the training set were set aside for validation when training the contrastive encoder.

3.6.1.4 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.[9] 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 has 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. Table 3.1 summaries the dataset statistics.

3.6.2 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 GDoS, the tag 'interj' was excluded as it is not present in the dataset.

[9] Code and data available at: https://github.com/zhewei-sun/slanggen

3.6.3 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 experimented with a combined prior where the two priors are combined using element-wise multiplication and re-normalization.

3.6.4 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 subword information on 600 billion tokens from Common Crawl[10] 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 Eq. 3.7. 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.

3.6.5 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 embed-

[10] http://commoncrawl.org

dings 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.

3.7 Experiments

3.7.1 Slang generation

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 going to predict the ground-truth slang recorded in the dictionary? Note that the goal here is not to reproduce the slang dictionary. The slang word recorded in a dictionary is among one of the plausible expressions that can extend to the slang sense. However, the recorded word choice has gained enough traction to enter a slang dictionary likely due to having high semantic plausibility compared to alternative words. Therefore, a good slang generation model should assign high probabilities to the ground-truth slang word choice.

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 3.3 as an illustration. The AUC metric is similar to and a

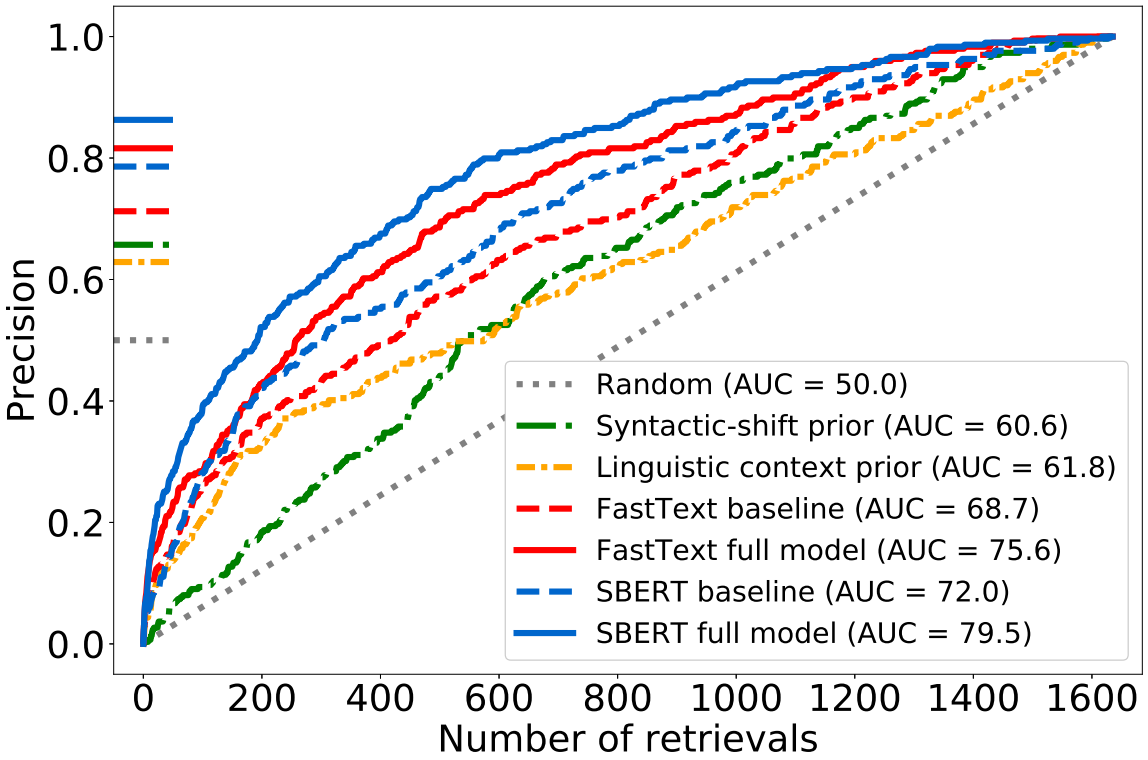


Figure 3.3: ROC curves for slang generation in OSD test set. Collaborative-filtering prototype model was used as the word choice model. Ticks on the y-axis indicate median precision of the models.

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 3.6.4). 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 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 3.2.

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.

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.

3.7.2 Evaluation on historic slang

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

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

predicted historical 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

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 3.3: 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 set sizes shown) and trained initially on 20,899 GDoS definitions prior to the 1960s. Test set entries from all previous decades are included in the training set.

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 given the existing slang usages at a specific time.

Table 3.3 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 the same predictive power over the emergence of future slang.

3.7.3 Zero-shot vs. few-shot generation

We find that the performance of our models vary substantially depending on 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.

(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 3.4: 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).

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 3.4 outlines the AUC scores of the collaboratively filtered prototype models under few-shot and zero-shot settings. For each dataset, we partitioned the corre-

sponding 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 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 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.

3.7.4 Synonymy in slang

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 3.4 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 3.5. 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, our full models show consistent improvement over the respective baselines across different degrees of synonymy, particularly with the SBERT based full model which offers substantial improvement in most cases.

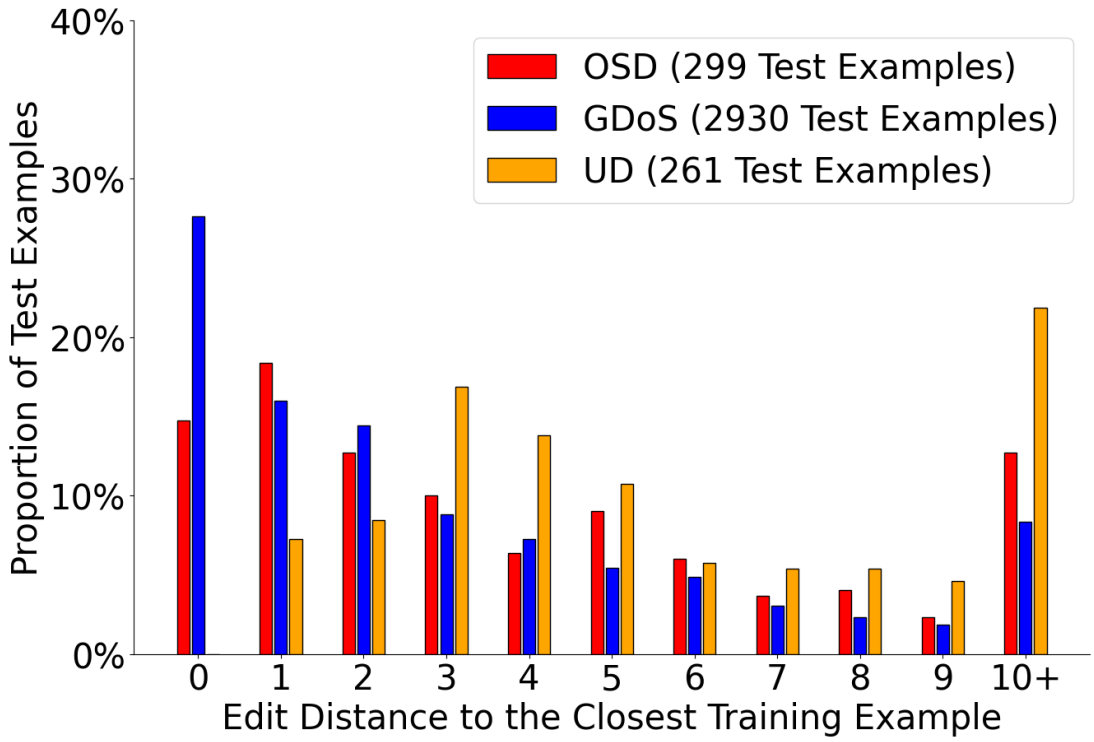


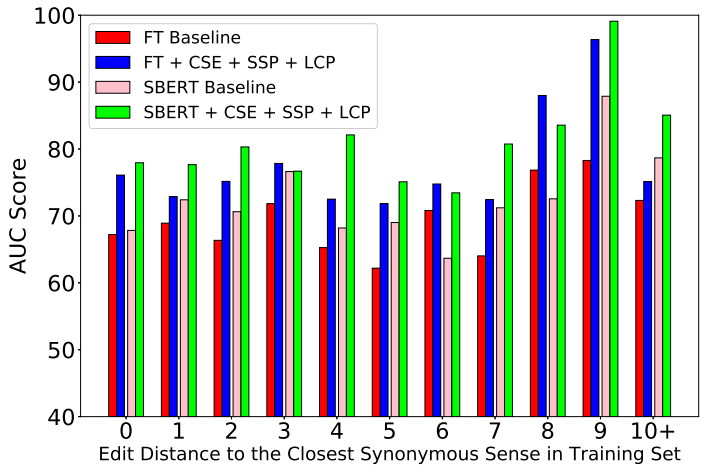
Figure 3.4: Degree of synonymy shared between the test examples and the respective training examples for each of the three datasets.

3.7.5 Comparing sense representations

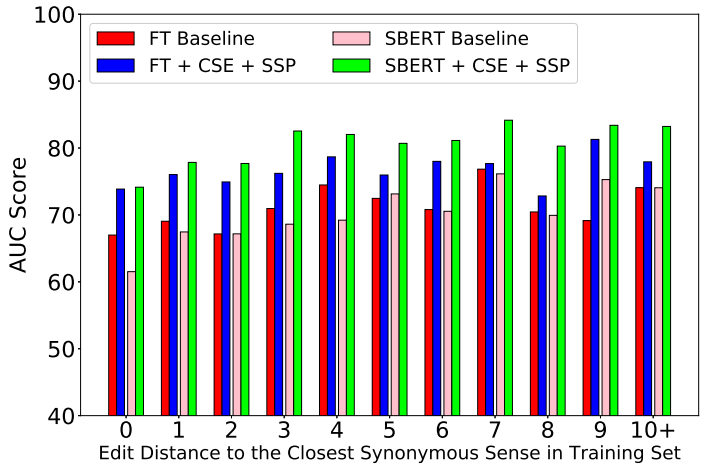
To understand the consequence of contrastive embedding, we compute the relative distances between conventional and slang senses of a word in an embedding space. This shows the extent to which the learned semantic relations may generalize. We measured the Euclidean distance between each slang embedding with the prototype sense vector of all candidate words, without applying the probabilistic choice models.

Table 3.5 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 embeddings of corresponding slang and conventional senses (from the same word), as indicated by lower mean ranks after the embedding procedure is applied. Under both fastText and SBERT, we obtained significant improvement (i.e. lower embedding distance rank) on both the OSD and GDoS test sets ($p < 0.001$). On UD, the improvement is significant for SBERT

(a) Online Slang Dictionary (OSD)



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



(c) Urban Dictionary (UD)

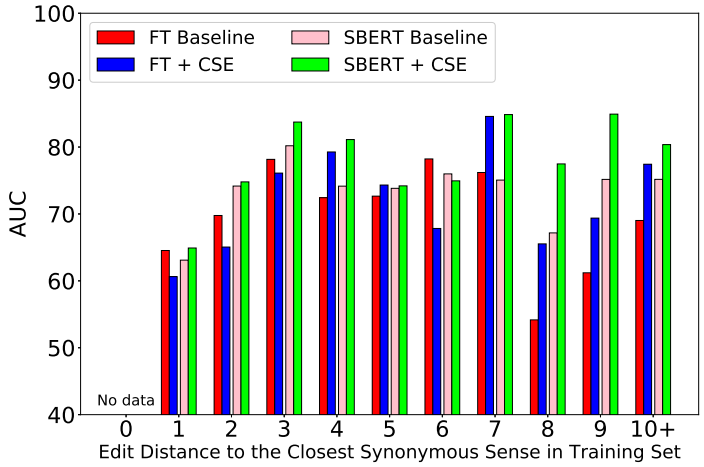


Figure 3.5: Model AUC scores (%) under test sets with different degrees of synonymy present in training, for the baselines and the best performing models (under collaborative-filtering prototype).

(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 3.5: Mean embedding distance ranks based on Euclidean distances from slang sense embeddings to prototypical conventional sense embeddings.

($p = 0.0062$) but marginal for fastText ($p = 0.098$).[11] The improved results on the test sets illustrate the ability of our constrative learning scheme to effectively generalize common slang semantic extension patterns from the training data.

3.7.6 Example generations

Table 3.6 shows 5 example slang usages 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.,

[11] The p-values were computed using the Wilcoxon signed-rank test on pairs of sense embedding distances before and after applying constrastive learning.

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 3.6: 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. SE is an abbreviation for 'Standard English' used in GDoS.

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.

3.8 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 the semantics of slang sense extension, and we have shown the promise of the learned semantic space for capturing semantic extension patterns for slang that are attested in slang dictionary data. We show in the next chapter how the generative semantic model can be applied to enhance more practical aspects of slang NLP, particularly the automated interpretation and translation of slang.

Chapter 4

Slang interpretation and translation

The contents of this chapter are based on my previous publication (Sun et al., 2022).

4.1 Motivation

Building on the model for slang generation, this chapter considers the inverse problem of slang interpretation that has more direct applications in natural language processing particularly machine translation (e.g., of informal language). We combine the generative semantic model of slang with context-based models in a semantically informed interpretation framework that infers the intended meaning of a target slang. We show that the generative semantic model can be applied to improve any context-based baselines without the need to perform task-specific fine-tuning.

The interpretation and translation of slang can be difficult for both humans and machines. Empirical studies have shown that, although it is done instinctively, interpretation and translation of unfamiliar or novel slang expressions can be quite hard for humans (Braun and Kitzinger, 2001; Mattiello, 2009). Similarly, these problems are also notoriously difficult for natural language processing (NLP) systems, which presents a critical challenge to downstream applications such as natural language

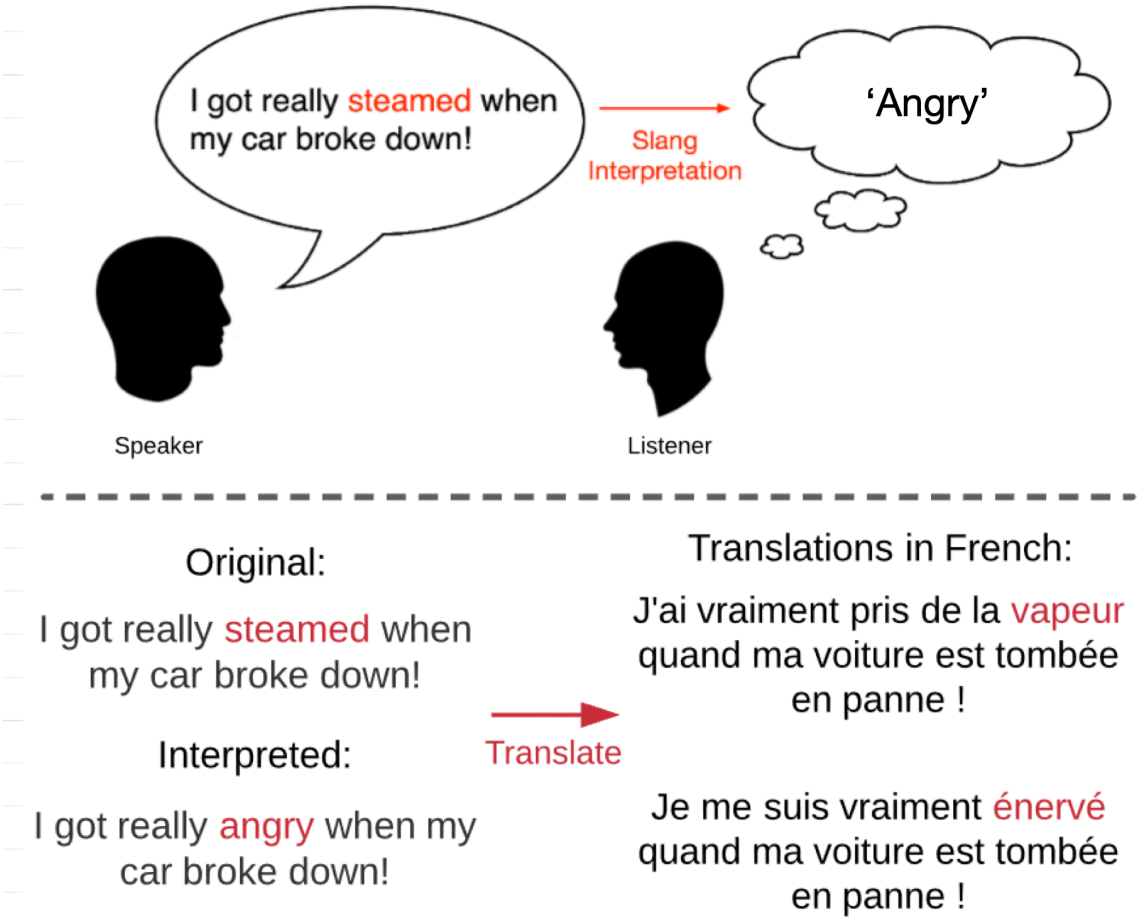


Figure 4.1: Illustrations of slang interpretation in English (top panel) and slang translation (bottom panel) from English to French on the original sentence (nonsensical), or on the interpreted version of the sentence (sensical).

understanding and machine translation.

Consider the sentence "I got really *steamed* when my car broke down". As illustrated in Figure 4.1, directly applying a translation system such as Google Translate on this raw English sentence would result in a nonsensical translation of the slang term *steamed* in French. This error is due partly to the underlying language model that fails to recognize the flexible extended use of the slang term from its conventional meaning (e.g., 'Vapor') to the slang meaning of 'Angry'. However, if knowledge about such semantic extensions can be incorporated into interpreting the slang prior to translation, as Figure 4.1 shows the system would be quite effective in translating the intended meaning.

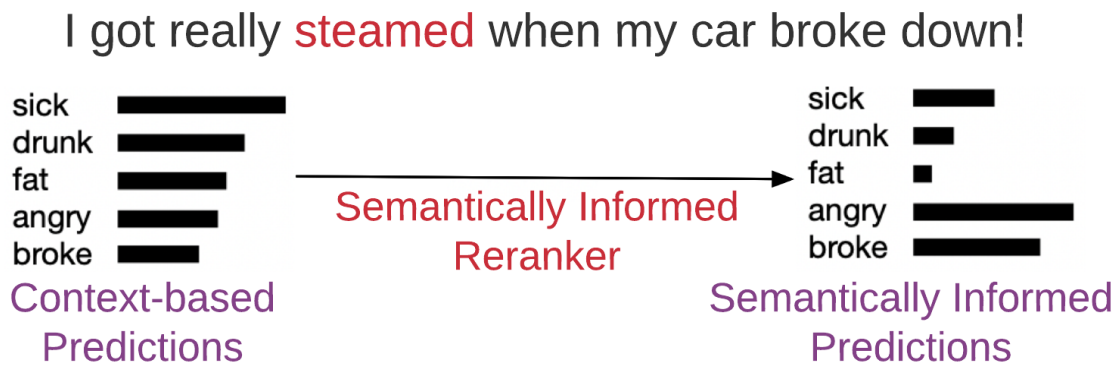


Figure 4.2: Overview of the slang interpretation method. A context-based model is first applied to obtain a list of candidate interpretations. Then, each candidate is evaluated by a generative model of slang semantics to output a reranked list of interpretations.

Here we consider the problem of slang interpretation illustrated in the top panel of Figure 4.1. Given a target slang term like *steamed* in a novel query sentence, we want to automatically infer its intended meaning in the form of a definition (e.g., 'Angry'). Tackling this problem has implications in both machine interpretation and understanding of informal language within individual languages and translation between languages.

One natural solution to this problem is to use contextual information to infer the meaning of a slang term. Figure 4.2 illustrates this idea by showing the top infilled words predicted under a GPT-2 (Radford et al., 2019) based language infill model (Donahue et al., 2020). Each of these words can be considered a candidate paraphrase for the target slang *steamed* conditioned on its surrounding words. Although the ground-truth meaning 'Angry' is among the list of top candidates, this model infers 'Sick' as the most probable interpretation. A similar context-based approach has been explored in a previous study led by Ni and Wang (2017) showing that a sequence-to-sequence model trained directly on a large number of pairs of slang-containing sentences along with their corresponding definitions from Urban Dictionary can be a useful starting point toward the automated interpretation of slang.

We present an alternative approach to slang interpretation that builds on but goes beyond the context-based models. Inspired by generative models of slang, we consider

slang interpretation to be the inverse process of slang generation and propose a semantically informed framework that takes into account both contextual information and knowledge about slang meaning extensions (e.g., 'Vapor'→'Angry') in inferring candidate interpretations. Our framework incorporates a semantic model of slang that uses contrastive learning to capture semantic extensions that link conventional and slang meanings of words. Under this framework, meanings that are otherwise far apart can be brought close, resulting in a semantic space that is sensitive to the flexible extended usages of slang. Rather than using this learned semantic space to generate novel slang usages, we apply it to the inverse problem of slang interpretation by checking whether a candidate interpretation may be suitably expressed as a slang using the to-be-interpreted slang expression. For example, 'Sick' and 'Angry' can both replace the slang *steamed* in a given context, but 'Angry' may be a more appropriate meaning to be expressed using *steamed* in the slang context. As such, we build a computational framework that takes into account the semantic knowledge of words as well as the context of slang in the interpretation process.

Following Ni and Wang (2017), we assume that the to-be-interpreted slang has already been detected in a sentence. In our experiments, example usage sentences in slang dictionaries are taken as slang-containing sentences. In practice, detection can be achieved by setting up a slang detection task directly as in work of Pei et al. (2019). Alternatively, novel sense identification methods (e.g., Lau et al., 2012; Cook et al., 2013) can be applied to find word usages attached with novel senses.[1]

Figure 4.2 shows an overview of our proposed method. We begin with a set of candidate interpretations informed by a context-based model (e.g., a language infill model), where the set would contain a list of possible meanings that fit reasonably in the given context. We then rerank this set of candidate interpretations by selecting the meaning that is most likely to be extended as slang from the to-be-interpreted slang expression.

[1]In this case, additional care must be taken to ensure that the detected sense indeed reflects slang usage, as novel conventional senses may also be detected

For the scope of this chapter, we focus on interpreting cases of reuse because such cases of slang cannot be readily addressed by existing dictionary-based approaches (e.g., Pal and Saha, 2013) or models of out-of-vocabulary words (e.g., Sennrich et al., 2016; Pinter et al., 2017). However, extensive studies in slang have suggested that a high proportion of slang usages relies on the extended reuse of existing word forms (Warren, 1992; Green, 2010; Eble, 2012). We show that our framework can enhance large language models in slang interpretation in English and slang translation from English to other languages.[2]

4.2 Problem formulation

We define slang interpretation formally as follows. Given a target slang term S in context C_S of a query sentence, interpret the meaning of S by a definition M.[3] The context is an important part of the problem formulation since a slang term S may be polysemous and context can be used to constrain the interpretation of its meaning. We define a slang interpreter I probabilistically as:

$$I(S, C_S) = \arg\max_M P(M|S, C_S) \tag{4.1}$$

Given this formulation, we retrieve an n-best list of candidate interpretations $\mathcal{K}$ (i.e., $|\mathcal{K}| = n$) based on an interpretation model of choice $P(M|S, C_S)$. Here, we consider two baseline models for $P(M|S, C_S)$: 1) a language-model (LM) based approach that treats slang interpretation as a cloze task, and 2) a sequence-to-sequence based approach similar to work by Ni and Wang (2017).

[2]Code and data available at: https://github.com/zhewei-sun/slanginterp
[3]In our experiments, M is a definition sentence.

4.3 Baseline approaches

4.3.1 Unsupervised language model based interpretation

The first model we consider is a language infill model in a cloze task, in which the model itself is based on large pre-trained language models such as GPT-2 (Radford et al., 2019). Although slang expressions may make sporadic appearances during training, this model is not trained specifically on a slang-related task and thus serves as a baseline that reflects the state-of-the-art language-model based NLP systems (e.g., Donahue et al., 2020).

Given context C_S containing target slang S, we blank out S in the context and ask the language infill model to infer the most likely words to fill in the blank. This results in a probability distribution $P(w|C_S\backslash S)$ over candidate words w. The infilled words can then be viewed as candidate interpretations of the slang S:

$$I(S, C_S) = D\left[\arg\max_w \left(P(w|C_S\backslash S) + \begin{cases} 1, & \text{if } T(w) = T(C_S\backslash S). \\ 0, & \text{otherwise.} \end{cases}\right)\right] \tag{4.2}$$

Here, D is a dictionary lookup function that maps a candidate word w to a definition sentence. In this case, we constrain the space of meanings considered to the set of all meanings corresponding to words in the lexicon.[4] Additionally, we apply a Part-of-Speech (POS) tagger T to check whether the candidate word w shares the same POS tag as the blanked-out word in the usage context. Words that share the same POS tags are preferred in the list of n-best retrievals.

This baseline approach by itself does not take into account any (semantic) information from the target slang S. In the case where two distinctive slang terms may be placed in the same context, the model would generate the exact same output. However, this LM based approach does not require task-specific data to train.

[4] In doing so, we assume that the intended meaning can be expressed using an existing word in our lexicon. Thus, a limitation of the LM-based approach is that it cannot accurately interpret novel concepts that have not entered the standard lexicon.

We show later that by reranking language model outputs, it is possible to achieve state-of-the-art performance using much less on-task data than existing approaches.

4.3.2 Supervised deep learning based interpretation

Ni and Wang (2017) partly addressed the context-only limitation by encoding the slang term using a character-level recurrent neural network in an end-to-end model inspired by the sequence-to-sequence architecture for neural machine translation (Sutskever et al., 2014). We implement their dual encoder architecture as an alternative context-based interpreter to LM. In this model, separate LSTM encoders are applied on the context C_S and the character encoding of the to-be-interpreted slang S respectively. The two encoders are then linearly combined using learned parameters:

$$h_{encode} = h_{context}W_{context} + h_{char}W_{char} + B \tag{4.3}$$

Here, $h_{context}$ and h_{char} are the final LSTM hidden states from the context encoder and character-level encoder respectively. $W_{context}$, W_{char}, and B are weights of the final linear layer to compute the combined state h_{encode}.

The combined state is passed onto an LSTM decoder to train against the corresponding definition sentence in Urban Dictionary (as in the original work of Ni and Wang 2017). For inference, beam search (Graves, 2012) is applied to decode an n-best list of candidate definition sentences:

$$I(S, C_S) = \arg\max_M DE(M|S, C_S) \tag{4.4}$$

Where DE is the dual encoder neural network.

4.4 Semantically-informed slang interpretation

4.4.1 Motivation

One key problem with the context-based approach is that it tends to rely on the contextual features surrounding the target slang but does not model flexible meaning extensions of the slang word itself. Similar issues are present in a language-model based approach, whereby one can use an infill model to infer the meaning of a target slang based solely on its surrounding words. Our work extends these context-based approaches by jointly considering the contextual and semantic appropriateness of a slang expression in a sentence, using generative semantic models of slang.

4.4.2 Generative model of slang semantics

From the baseline models, we obtain an n-best list of candidate interpretations $\mathcal{K}$ for the target slang S in context C_S. Given this list, we wish to model the semantic plausibility of each candidate interpretation $k \in \mathcal{K}$. Specifically, we ask how likely one would relate the (conventional meaning of) target slang expression S to a candidate interpretation k. Similar to slang generation, we model the relationship between a to-be-expressed meaning and a word form using the prototype model (Rosch, 1975; Snell et al., 2017). We adapt this model in the context of slang interpretation:

$$\begin{aligned} f(k, S) &= sim(E_k, E_S) \\ &= \exp\Big(- \frac{d(E_k, E_S)}{h_m} \Big) \end{aligned} \tag{4.5}$$

E_k is an embedding for a candidate interpretation k and E_S is the prototypical conventional meaning of S computed by averaging the embeddings of its conventional meanings in dictionary ($\mathcal{E}_S$):

$$E_S = \frac{1}{|\mathcal{E}_S|} \sum_{E_{S_i} \in \mathcal{E}_S} E_{S_i} \tag{4.6}$$

The similarity function f can then be computed by taking the negative exponential of the Euclidean distance (d) between the two resulting semantic embeddings. h_m is a kernel width hyperparameter.

We learn semantic embeddings E_k and E_{S_i} under a max-margin triplet loss scheme, where embeddings of slang sense definitions (E_{SL}) are brought close in Euclidean space to those of their conventional sense definitions (E_P) yet kept apart from irrelevant word senses (E_N) by a pre-specified margin m:

$$Loss = \Big[d(E_{SL}, E_P) - d(E_{SL}, E_N) + m\Big]_+ \tag{4.7}$$

The resulting contrasive sense encodings are shown to be sensitive to slang semantic extensions that have been observed during training. We leverage this knowledge to check whether pairing a candidate interpretation k with the slang expression S is likely given the common semantic extensions observed in slang usages. The resulting scores can then be used to rerank the candidate interpretations.

4.4.3 Semanticically-informed reranking

We define a semantic scorer g over the set of candidate interpretations $\mathcal{K}$ and the to-be-interpreted slang S. The candidates are reranked based on the resulting scores to obtain semantically informed slang interpretations (SSI):

$$SSI(\mathcal{K}) = \arg\max g(k, S) \tag{4.8}$$

We define $g(\mathcal{K}, S)$ as a score distribution over the set of candidates $\mathcal{K}$ given slang S, where each score is computed by checking the semantic appropriateness of a candidate meaning $k \in \mathcal{K}$ with respect to target slang S by querying the semantic model f from Equation 4.5:

$$g(k, S) = P(k|S) \propto f(k, S) \tag{4.9}$$

Dataset	# of unique slang word forms	# of slang definition entries	# of context sentences	# of definitions in the test set	# of context sentences in the test set
OSD	1,635	2,979	3,718	299	405
UD	9,474	65,478	65,478	1,242	1,242

Table 4.1: Summary of dataset statistics for the online slang dictionaries used in the slang interpretation study.

In addition, we apply collaborative filtering (Goldberg et al., 1992) to account for a small neighborhood of words $L(S)$ akin to the slang expression S in conventional meaning:

$$g^*(k, S) \propto \sum_{S' \in L(S)} sim(S, S') g(k, S') \tag{4.10}$$

$$sim(S, S') = \exp\Big(- \frac{d(S, S')}{h_{cf}} \Big) \tag{4.11}$$

Here, $d(S, S')$ is the cosine distance between the word vectors of two slang expressions and h_{cf} is a hyperparameter controlling the kernel width. The collaborative filtering step encodes intuition from studies in historic semantic change that similar words tend to extend to express similar meanings (Lehrer, 1985; Xu and Kemp, 2015), which was found to extend well in the case of slang (see Chapter 3).

4.5 Experimental setup

4.5.1 Datasets

We use two online English slang dictionary resources to train and evaluate our proposed slang interpretation framework: 1) the Online Slang Dictionary (OSD)[5] dataset from Chapter 3 and 2) a collection of Urban Dictionary (UD)[6] entries from 1999 to 2014 collected by Ni and Wang (2017). Each dataset contains slang gloss entries including a slang's word form, its definition, and at least one corresponding exam-

[5]OSD: http://onlineslangdictionary.com
[6]UD: https://www.urbandictionary.com

ple sentence containing the slang term. We use the same training and testing split provided by the original authors and only use entries where a corresponding non-informal entry can be found in the online version of the Oxford Dictionary (OD) for English,[7] which allows the retrieval of conventional senses for all slang expressions considered. We also filter out entries where the example usage sentence contains none or more than one occurrences of the corresponding slang expression. When a definition entry has multiple example usage sentences, we treat each example sentence as a separate data entry, but all data entries corresponding to the same definition entry will only appear in the same data split. Table 4.1 shows the size of the datasets after pre-processing. While OSD contains higher quality entries, UD offers a much larger dataset. We thus use OSD to evaluate model performance in a low resource scenario and UD for evaluation of larger neural network based approaches.

4.5.2 Training procedures

4.5.2.1 Baseline Models

We train two context-based slang interpreters described in Section 4.3 as our baseline models. For the LM-based interpreter, we use a pre-trained language infill model from Donahue et al. (2020) based on the GPT-2 (Radford et al., 2019) architecture. Here, we obtain the n-best list of interpretations by retrieving the list of infilled words with the highest infill probability. Words containing non-alphanumeric characters are filtered out. For the dictionary lookup function D in Equation 4.2, if a matching dictionary entry can be found in Oxford Dictionary (OD), the first definition sentence is retrieved as the definition sentence for the input word. Otherwise, the word itself is used as the definition. In addition to the word's original form, we apply lemmatization or stemming to the original form using NLTK (Bird et al., 2009) to find matching dictionary entries. To check for Part-of-Speech (POS) tags, we apply the Flair tagger (Akbik et al., 2018) on the context sentence with the slang expression

[7] OD: https://en.oxforddictionaries.com

replaced by a mask token and use counts from Histwords (Hamilton et al., 2016) to determine POS tags for individual words.

To train the Dual Encoder, we use LSTM encoders with 256 and 1024 hidden units to encode a slang expression's spelling and its usage context respectively, with 100 and 300 dimensional input embeddings for the characters and words respectively. Following Ni and Wang (2017), we use random initialization for the input embeddings and use stochastic gradient descent (SGD) with an adaptive learning rate. We train the model for 20 epochs beginning with a learning rate of 0.1 and add an exponential decay of 0.9 every epoch. We reserve 5% of the training examples as a development set for hyperparameter tuning. We train the model for 20 epochs on a Nvidia Titan V GPU and takes 12 hours to complete. During inference, we obtain the n-best list of interpretations by running a beam search of corresponding beam width on the LSTM decoder.

4.5.2.2 Semantic Reranker

We obtain the contrastive sense encodings (CSE) described in Section 4.4.2 by using 768-dimensional Sentence-BERT (Reimers and Gurevych, 2019) embeddings as our baseline embedding. We train the contrastive network with a 1.0 margin (m in Equation 4.7) using Adam (Kingma and Ba, 2015) with a learning rate of 2^{-5}, resulting in 768-dimensional definition sense representations. We reserve 5% of the training examples as a development set for hyperparameter tuning. The contrastive models are trained on a Nvidia Titan V GPU for 4 epochs. The OSD model took 85 minutes to train and the UD model took 8 hours. We follow the training procedure from Chapter 3 to estimate the kernel width parameters (h_m in Equation 4.5 and h_{cf} in Equation 4.11) via generative training when it is computationally feasible to do so and otherwise use 0.1 as our default value.

We check the similarity between two expressions in Equation 4.11 by comparing their fastText (Bojanowski et al., 2017) embeddings. For collaborative filtering, the

neighborhood of words $L(S)$ in Equation 4.10 is defined as the 5 closest words (including the query word itself) in the dataset's slang expression vocabulary to the query word, measured in terms of cosine similarity between their respective fastText embeddings. We use the list of stopwords from NLTK (Bird et al., 2009) to check whether a word is a content word. We apply the *simple_preprocess* routine from Gensim (Rehurek and Sojka, 2011) before checking for the degree of content word overlap between two sentences.

4.5.3 Evaluation methods

We evaluate the semantically informed and baseline interpretation models in a multiple choice task. In this task, each query is paired with a set of candidate definitions for the target slang in the query. One of these definitions is the ground-truth meaning of the target slang, while the other definitions are incorrect, i.e., negative entries sampled from the training set that are all taken from the corresponding slang dictionary. To score a model, each definition sentence is first compared with the model-predicted definition by computing the Euclidean distance between their respective Sentence-BERT (Reimers and Gurevych, 2019) embeddings. The ideal model should produce a definition that is semantically closer to the ground-truth definition, more so than the other competing negatives. For each dataset, we sample two sets of negatives. The first set of negative candidates contains only definition sentences from the training set that are distinct from the ground-truth definition. We consider two definition sentences to be distinct if the overlap in the number of content words is less than 50%. The other set of negative definitions is sampled randomly. We measure the performance of the models by computing the standard mean reciprocal rank (MRR) of the ground-truth definition's rank when checked against 4 other sampled negative definitions.

We train the semantic reranker on all definition entries in the respective training sets from the two data resources. When training the Dual Encoder, we use 400,431

Model	Distinctively sampled candidates	Randomly sampled candidates
Dataset 1: Online Slang Dictionary (OSD)		
Language Infill Model (LM Infill) (Donahue et al., 2020), $n = 50$	0.532	0.502
+ Semantically Informed Slang Interpretation (SSI)	**0.557**	**0.563**
Dual Encoder* (Ni and Wang, 2017), $n = 5$	0.584	0.583
+ SSI	**0.592**	**0.588**
Dual Encoder*, $n = 50$	0.568	0.602
+ SSI	**0.616**	**0.607**
* Dual Encoders trained on UD data after filtering out slang in OSD test set.		
Dataset 2: Urban Dictionary (UD) (Ni and Wang, 2017)		
LM Infill, $n = 50$	0.517	0.521
+ SSI	**0.569**	**0.579**
Dual Encoder, $n = 5$	0.556	0.555
+ SSI	**0.573**	**0.572**
Dual Encoder, $n = 50$	0.547	0.550
+ SSI	**0.582**	**0.584**

Table 4.2: Evaluation of English slang interpretation measured in mean-reciprocal rank (MRR). Predictions are ranked against 4 negative candidates distinctively or randomly sampled, yielding an MRR of 0.457 for the random baseline.

out-of-vocabulary slang entries (i.e., entries with a slang expression that does not contain a corresponding lexical entry in the standard dictionary) from UD in addition to the in-vocabulary entries used to train the reranker. This is necessary since the baseline Dual Encoder performs poorly without a large number of training entries. Similarly, training the Dual Encoder directly on the OSD training set does not result in an adequate model for comparison. We instead train the Dual Encoder on all UD entries and experiment with the resulting interpreter on OSD. Any UD entries corresponding to words found in the OSD testset are filtered out in this particular experiment.

4.6 Experimental results

4.6.1 Slang interpretation

Table 4.2 summarizes the multiple-choice evaluation results on both slang datasets. In all cases, applying the semantically informed slang interpretation framework improves

the MRR of the respective baselines under both types of negative candidate sampling. On the UD evaluation, even though the language infill model (LM Infill) is not trained on this specific task, LM infill based SSI is able to select better and more appropriate interpretations than the dual encoder baseline, which is trained specifically on slang interpretation with more than 7 times the number of definition entries for training. We also find that while increasing the beam size (specified by n) in the sequence-to-sequence based Dual Encoder model impairs its performance, SSI can take advantage of the additional variation in the generated candidates and outperform its counterpart with a smaller beam size.

[Example 1]	
Query (target slang in ***bold italic***):	That chick is ***lit***!
Groundtruth definition of target slang:	Attractive.
LM Infill baseline prediction:	Cute, beautiful, adorable.
LM Infill + SSI prediction:	Hot, cool, fat.
Dual Encoder baseline prediction:	Another word for bitch.
Dual Encoder + SSI prediction:	Word used to describe someone who is very attractive.
[Example 2]	
Query:	That Louis Vuitton purse is ***lush***!
Groundtruth definition of target slang:	High quality, luxurious. (British slang.)
LM Infill baseline prediction:	Amazing, beautiful, unique.
LM Infill + SSI prediction:	Lovely, stunning, expensive.
Dual Encoder baseline prediction:	Something that is cool or awesome.
Dual Encoder + SSI prediction:	An adjective used to describe something that is not cool.
[Example 3]	
Query (target slang in ***bold italic***):	That girl has a ***donkey***.
Ground-truth definition of target slang:	Used to describe a girl's butt in a good way.
LM Infill baseline prediction:	Name, crush, boyfriend.
LM Infill + SSI prediction:	Horse, dog, puppy.
Dual Encoder baseline prediction:	Penis.
Dual Encoder + SSI prediction:	Girl with big ass and big boobs.
[Example 4]	
Query:	I am an ***onion***.
Ground-truth definition of target slang:	A native of Bermuda.
LM Infill baseline prediction:	Adult, man, athlete.
LM Infill + SSI prediction:	Ren, adult, guard.
Dual Encoder baseline prediction:	An idiot.
Dual Encoder + SSI prediction:	An asian person.
[Example 5]	
Query:	In Blastem version 4, they really ***nerf*** the EnemyToaster.
Ground-truth definition of target slang:	In an update or sequel to a video game, to make a weapon weak or weaker, such that it's like a Nerf gun.
LM Infill baseline prediction:	Were, called, attack.
LM Infill + SSI prediction:	Made, hacked, came.
Dual Encoder baseline prediction:	To do something.
Dual Encoder + SSI prediction:	To beat someone in the face with your penis.

Table 4.3: Example OSD slang entries with predicted definitions from both the language infill model (LM Infill) and the Dual Encoder model with $n = 50$, along with predictions from the corresponding semantically informed slang interpretation (SSI) models.

[Example 6]	
Query:	I heard Steve was sent to the ***cooler*** for breaking and entering.
Ground-truth definition of target slang:	Reform school.
LM Infill baseline prediction:	School, house, class.
LM Infill + SSI prediction:	Bathroom, kitchen, grounds.
Dual Encoder baseline prediction:	Slang term for the police.
Dual Encoder + SSI prediction:	One of the most dangerous things in the world the best.
[Example 7]	
Query:	Do you have any ***safety***
Ground-truth definition of target slang:	Marijuana.
LM Infill baseline prediction:	Money, friends, cash.
LM Infill + SSI prediction:	Self, shoes, money.
Dual Encoder baseline prediction:	Marijuana.
Dual Encoder + SSI prediction:	Word that is used to describe something that is very good.

Table 4.4: Continuation of Table 4.3 showing additional interpretation examples.

Table 4.3 and Table 4.4 provides example interpretations predicted by the models. The *lit* example shows a case where the semantically informed models were able to correctly pinpoint the intended definition, among alternative definitions that describe individuals. The *lush* example suggests that the SSI model is not perfect and points to common errors made by the model including predicting definitions that are more general and applying incorrect semantic extensions. In this case, the model predicts the slang *lush* to mean 'Something that is not cool' because polarity shift is a common pattern in slang usage (Eble, 2012), even though the ground-truth definition does not make such a polarity shift in this specific example.

Note that the improvement brought by SSI is less prominent in the OSD experiment where the Dual Encoder trained on UD was used. This is expected because the Dual Encoder is trained to generate definition sentences in the style of UD entries, whereas the SSI is trained on OSD definition sentences instead. The mismatch in style between the two datasets might have caused the difference in performance gain.

Model	Distinct negatives	Random negatives
LM Zero-shot, $n = 50$	0.444	0.443
+ SSI	**0.571**	**0.565**
LM Few-shot, $n = 50$	0.504	0.513
+ SSI	**0.567**	**0.564**

Table 4.5: Interpretation results on OSD measured in mean-reciprocal rank (MRR) before and after finetuning the language infill model.

4.6.2 Few-shot slang interpretation

Recent studies in deep learning have shown that large neural network based models such as GPT-3 excel at learning new tasks in a few-shot learning setting (Brown et al., 2020). We examine to what extent the superior performance of our SSI framework may be affected by fine-tuning the LM baseline model in zero-shot and few-shot scenarios. We fine-tune the language infill model (LM Infill) on the first example usage sentence that corresponds to each definition entry in the OSD dataset, resulting in 2,979 sentences. Given an example sentence, we mask out the slang expression and train the language infill model to predict the corresponding slang term. We randomly shuffle all examples and fine-tune LM Infill for one epoch. We then compare the resulting model with the off-the-shelf LM using examples in the test set that were not used in finetuning (i.e., entries with usage sentences that do not correspond to the first example usage sentence of a definition entry). This results in 106 novel examples for evaluation.

Table 4.5 shows the result of this experiment. While finetuning does improve test performance (a 6 point gain in MRR), it remains beneficial to consider semantic information in slang context. In both the zero-shot and the few-shot cases, SSI brings significant performance gain even though SSI itself has not seen examples from the test set during its training.

Model	Distinct negatives	Random negatives
Dual Encoder, $n = 5$	0.604	0.598
+ SSI	**0.612**	**0.599**
Dual Encoder, $n = 50$	0.583	0.570
+ SSI	**0.627**	**0.633**

Table 4.6: Interpretation results on OSD measured in mean-reciprocal rank (MRR) when training the Dual Encoder without filtering out entries corresponding to words in the OSD testset.

4.6.3 Effect of Context Length

In the model evaluation described in Section 4.6.1, we control for the content-word length of the usage context sentence to examine its effect with respect to interpretation performance for both the baseline and the semantically informed models. Figure 4.3 shows the results partitioned by the number of content words in the example usage sentence excluding the slang expression, evaluated against four distinctively sampled candidates. To our surprise, we do not observe any consistent trends when controlling for context length. Interpretation performance for both the context-based baseline models and their semantically informed variants is fairly consistent under different context length.

4.6.4 Finetuning Dual Encoder

We consider the case of finetuning the Dual Encoder by training it on all available UD data entries and test on the full OSD test set. Under this scenario, the Dual Encoder model would have seen examples of slang in the OSD test set, though the difference between the definition sentences and usage examples would not allow it to memorize the exact answer. While examining how much knowledge can be transfered from one dataset to another, we also apply the SSI reranker trained on OSD training data on the fine-tuned results to simulate a stronger baseline model. Table 4.6 shows the results. When compared to the zero-shot results in Table 4.2, finetuning on entries corresponding to the same slang, albeit coming from two very different resources, does noticeably improve interpretation accuracy. Moreover, applying SSI to the im-

proved interpretation candidates from the fine-tuned Dual Encoder further increases interpretation accuracy. This finding suggests that the improvement brought by SSI can indeed generalize in cases where the baseline context-based interpretation model outputs better interpretation candidates.

4.7 Application in slang translation

We next apply the slang interpretation framework to neural machine translation. Existing machine translation systems have difficulty in translating source sentences containing slang usage partly because they lack the ability to properly decode the intended slang meaning. We make a first attempt in addressing this problem by exploring whether machine interpretation of slang can lead to better translation of slang. Given a source English sentence containing a slang expression S, we apply the LM based slang interpreters to generate a paraphrased word to replace S. The paraphrased sentence would then contain the intended meaning of the slang in its literal form. Here, we take advantage of the LM-based approaches' ability to directly generate a paraphrase instead of a definition sentence (i.e., without dictionary lookup D in Equation 4.2), which allows direct insertion of the resulting interpretation into the original sentence.

4.7.1 Experimental setup

We perform our experiment on the OSD test set because it contains higher quality example sentences than UD. To mitigate potential biases, we consider only entries that correspond to single-word slang expressions, and that the slang has not been seen during training (where the slang attaches to a different slang meaning than the one in the test set). For the remaining 102 test entries, we obtain ground-truth translations by first manually replacing the slang word in the example sentence with its intended definition, condensed to a word or short phrase to fit into the context sentence. We

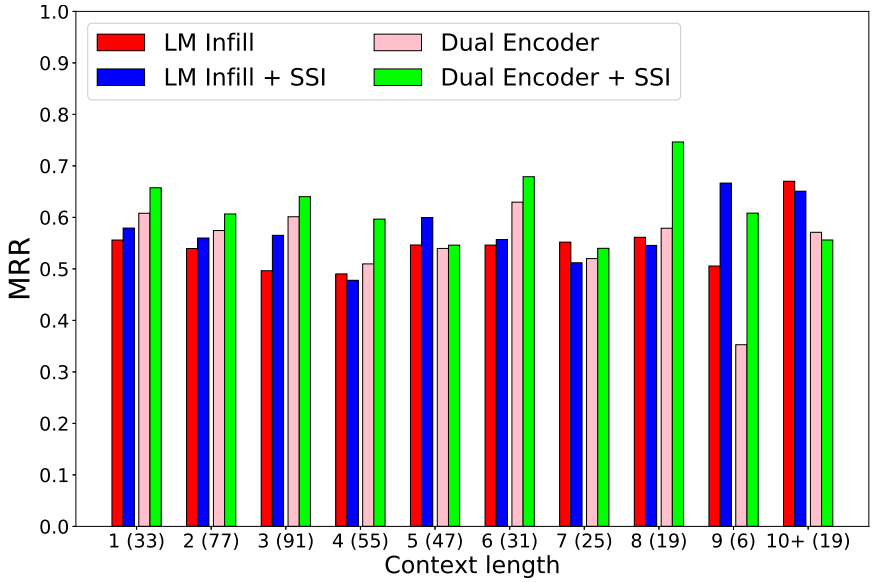
then translate the sentences to French and German using machine translation.

We make all machine translations using pre-trained 6-layer transformer networks (Vaswani et al., 2017) from MarianMT (Tiedemann and Thottingal, 2020), which are trained on a collection of web-based texts in the OPUS dataset (Tiedemann, 2012). Here, we select models pre-trained on web-based texts to maximize the baseline model's ability to correctly process slang. We evaluate the translated sentences using three metrics: 1) Sentence-level BLEU scores (Papineni et al., 2002) computed using *sentence_bleu* implementation from NLTK (Bird et al., 2009) with smoothing (*method4* in NLTK, Chen and Cherry, 2014) to account for sparse n-gram overlaps; 2) BLEURT scores (Sellam et al., 2020) computed using the pre-trained *BLEURT-20* checkpoint; 3) COMET scores (Rei et al., 2020) computed using the pre-trained *wmt20-comet-da* checkpoint. For COMET scores, we replace slang expressions in the source sentences with their literal equivalents to reduce confusion that the COMET model might have on slang. This is revelant only to COMET because it takes the source sentence as part of its evaluation, whereas the BLEU metrics only consider the ground-truth target sentence.

[Example 1]	
Query (target slang in ***bold italic***):	Let's smoke a ***bowl*** of marijuana.
Definition of target slang:	a marijuana smoking pipe. Most frequently bowls are made out of blown glass, but can be made of metal, wood, etc.
Ground-truth interpreted sentence:	Let's smoke a *pipe* of marijuana.
Original query sentence translation:	Faisons fumer un bol de marijuana. (*BLEU:* 78.1, *BLEURT:* 66.1, *COMET:* 1.05)
Gold-standard translation:	Faisons fumer une pipe de marijuana.
LM Infill interpretation & translation:	
(1) Let's smoke a *for* of marijuana.	Fumons un *pour* de la marijuana. (*BLEU:* 47.1, *BLEURT:* 20.6, *COMET:* -0.58)
(2) Let's smoke a *in* of marijuana.	On fume un *peu* (little) de marijuana. (*BLEU:* 51.6, *BLEURT:* 64.8, *COMET:* 0.48)
(3) Let's smoke a *myself* of marijuana.	Nous allons fumer *moi-même* de la marijuana. (*BLEU:* 51.8, *BLEURT:* 32.4, *COMET:* -0.55)
(4) Let's smoke a *or* of marijuana.	Fumons un *ou* de marijuana. (*BLEU:* 45.4, *BLEURT:* 32.2, *COMET:* -1.04)
(5) Let's smoke a *vapor* of marijuana.	Fumons une *vapeur* de marijuana. (*BLEU:* 56.4, *BLEURT:* 57.0, *COMET:* 0.40)
LM Infill + SSI interpretation & translation:	
(1) Let's smoke a *pot* of marijuana.	Faisons fumer un *pot* de marijuana. (*BLEU:* 79.5, *BLEURT:* 78.8, *COMET:* 1.15)
(2) Let's smoke a *pipe* of marijuana.	Faisons fumer une *pipe* de marijuana. (*BLEU:* 100.0, *BLEURT:* 99.1, *COMET:* 1.32)
(3) Let's smoke a *pack* of marijuana.	Faisons fumer un *paquet* de marijuana. (*BLEU:* 77.7, *BLEURT:* 68.3, *COMET:* 0.80)
(4) Let's smoke a *leaf* of marijuana.	Faisons fumer une *feuille* de marijuana. (*BLEU:* 79.9, *BLEURT:* 48.2, *COMET:* 1.21)
(5) Let's smoke a *cigarette* of marijuana.	Faisons fumer une *cigarette* de marijuana. (*BLEU:* 75.7, *BLEURT:* 81.7, *COMET:* 1.25)

Table 4.7: Examples of machine translation of slang, without or with the application of the SSI framework. The top 5 interpreted and translated sentences are shown for each model with BLEU, BLEURT, and COMET scores against the gold-standard translation shown in parentheses.

(a) OSD



(b) UD

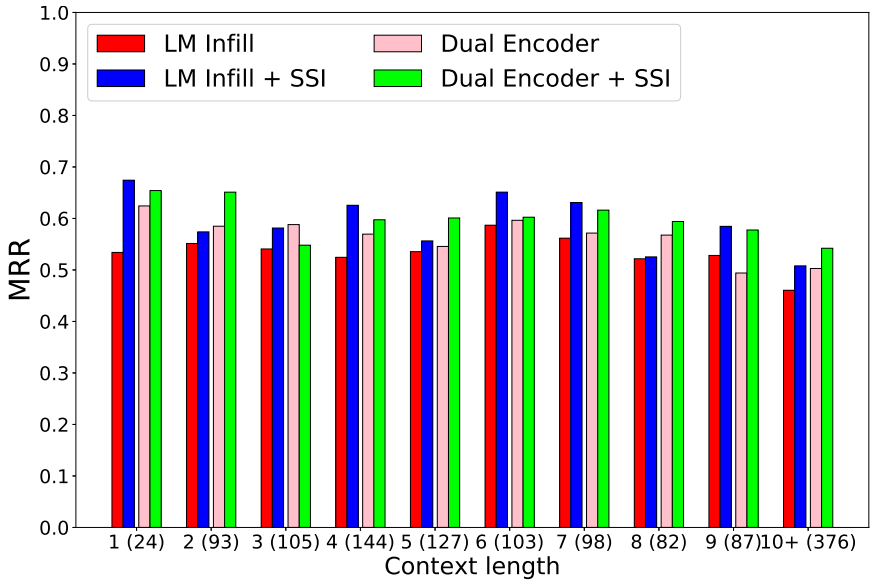
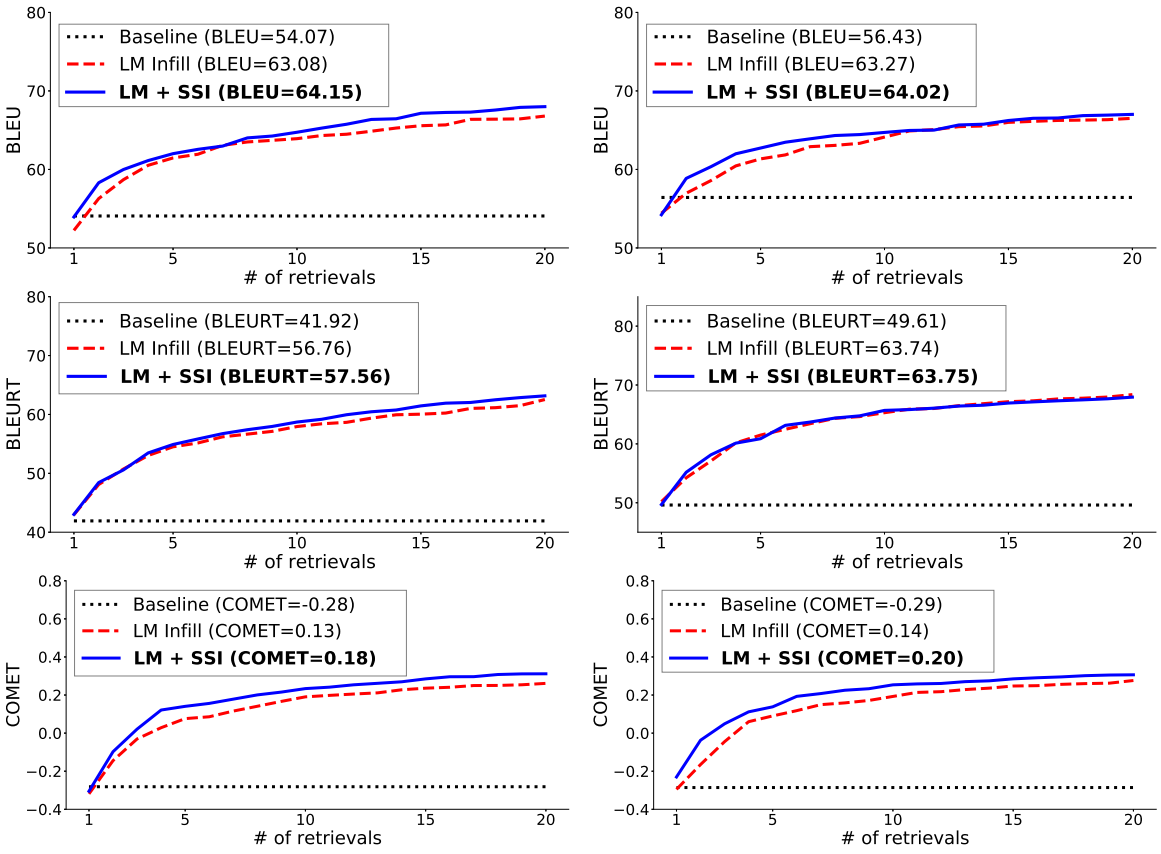


Figure 4.3: Evaluation of slang interpretation performance measured in mean-reciprocal rank (MRR) for all models with $n = 50$. Test entries are partitioned based on the number of content words (excluding the slang expression itself) found within the corresponding example usage sentence. Number of entries corresponding to each context length is shown in parenthesis on the x-axis legend.



(a) English to French (b) English to German

Figure 4.4: Translation scores of translated sentences with the slang replaced by n-best interpretations. Curves show sentence-level BLEU, BLEURT, and COMET scores of the best translation within the top-n retrievals. Aggregate scores integrated over the first 20 retrievals are shown in parenthesis. Baselines are obtained by directly translating the original sentence containing slang.

[Example 2]	
Query:	That band was so totally ***vast***.
Definition of target slang:	Cool or anything good.
Ground-truth interpreted sentence:	That band was so totally *cool*.
Original query sentence translation:	Ce groupe était si vaste. (*BLEU:* 53.2, *BLEURT:* 32.9, *COMET:* -0.59)
Gold-standard translation:	Ce groupe était tellement cool.
LM Infill interpretation & translation:	
(1) That band was so totally *popular*.	Ce groupe était tellement *populaire*. (*BLEU:* 74.5, *BLEURT:* 78.7, *COMET:* 0.43)
(2) That band was so totally *good*.	Ce groupe était si *bon*. (*BLEU:* 51.8, *BLEURT:* 77.0, *COMET:* 0.32)
(3) That band was so totally *different*.	Ce groupe était complètement *différent*. (*BLEU:* 57.2, *BLEURT:* 50.3, *COMET:* -0.07)
(4) That band was so totally *famous*.	Ce groupe était si *célèbre*. (*BLEU:* 54.4, *BLEURT:* 66.2, *COMET:* -0.21)
(5) That band was so totally *new*.	Ce groupe était totalement *nouveau*. (*BLEU:* 64.2, *BLEURT:* 50.2, *COMET:* -0.21)
LM Infill + SSI interpretation & translation:	
(1) That band was so totally *huge*.	Ce groupe était tellement *énorme*. (*BLEU:* 81.1, *BLEURT:* 56.0, *COMET:* 0.15)
(2) That band was so totally *big*.	Ce groupe était tellement *grand*. (*BLEU:* 83.0, *BLEURT:* 50.7, *COMET:* -0.19)
(3) That band was so totally *important*.	Ce groupe était si *important*. (*BLEU:* 55.9, *BLEURT:* 49.9, *COMET:* -0.58)
(4) That band was so totally *cool*.	Ce groupe était tellement *cool*. (*BLEU:* 100.0, *BLEURT:* 97.9, *COMET:* 1.29)
(5) That band was so totally *bad*.	Ce groupe était si *mauvais*. (*BLEU:* 52.3, *BLEURT:* 62.9, *COMET:* -0.48)

Table 4.8: Continuation of Table 4.7. Examples of machine translation of slang.

[Example 3]	
Query (target slang in ***bold italic***):	Man, I ain't been to that place in a ***fortnight***!
Definition of target slang:	An unspecific, but long-ish length of time.
Ground-truth interpreted sentence:	Man, I ain't been to that place in a *long time*!
Original query sentence translation:	Je ne suis pas allé à cet endroit en une quinzaine! (*BLEU:* 36.1, *BLEURT:* 61.2, *COMET:* 0.57)
Gold-standard translation:	Je n'y suis pas allé depuis longtemps!
LM Infill interpretation & translation:	
(1) Man, I ain't been to that place in a *while*!	Je ne suis pas allé à cet endroit depuis un *moment*! (*BLEU:* 46.9, *BLEURT:* 76.5, *COMET:* 0.88)
(2) Man, I ain't been to that place in a *million*!	Je ne suis pas allé à cet endroit dans un *million*! (*BLEU:* 38.8, *BLEURT:* 25.1, *COMET:* -1.17)
(3) Man, I ain't been to that place in a *both*!	Je ne suis pas allé à cet endroit dans les *deux*! (*BLEU:* 42.2, *BLEURT:* 25.7, *COMET:* -0.98)
(4) Man, I ain't been to that place in a *vanilla*!	Mec, je n'ai pas été à cet endroit dans une *vanille*! (*BLEU:* 16.2, *BLEURT:* 7.3, *COMET:* 1.53)
(5) Man, I ain't been to that place in a *ignment*!	Mec, je n'ai pas été à cet endroit dans un *ignement*! (*BLEU:* 16.2, *BLEURT:* 12.7, *COMET:* -1.31)
LM Infill + SSI interpretation & translation:	
(1) Man, I ain't been to that place in a *week*!	Je ne suis pas allé à cet endroit en une *semaine*! (*BLEU:* 38.2, *BLEURT:* 49.8, *COMET:* 0.45)
(2) Man, I ain't been to that place in a *minute*!	Je ne suis pas allé à cet endroit en une *minute*! (*BLEU:* 38.8, *BLEURT:* 42.5, *COMET:* -0.36)
(3) Man, I ain't been to that place in a *hour*!	Je ne suis pas allé à cet endroit en une *heure*! (*BLEU:* 38.7, *BLEURT:* 35.8, *COMET:* -0.51)
(4) Man, I ain't been to that place in a *decade*!	Je n'y suis pas allé depuis une *décennie* (*BLEU:* 68.8, *BLEURT:* 81.8, *COMET:* 1.03)
(5) Man, I ain't been to that place in a *day*!	Je ne suis pas allé à cet endroit en une *journée*! (*BLEU:* 37.1, *BLEURT:* 49.7, *COMET:* -0.30)

Table 4.9: Continuation of Table 4.8. Examples of machine translation of slang.

[Example 4]	
Query:	I want to go get coffee but it's ***bitter*** outside.
Definition of target slang:	Abbreviated form of bitterly cold.
Ground-truth interpreted sentence:	I want to go get coffee but it's *bitterly cold* outside.
Original query sentence translation:	Je veux aller prendre un café mais c'est amer dehors. (*BLEU:* 65.0, *BLEURT:* 59.8, *COMET:* 0.77)
Gold-standard translation:	Je veux aller prendre un café, mais il fait très froid dehors.
LM Infill interpretation & translation:	
(1) I want to go get coffee but it's *raining* outside.	Je veux aller prendre un café mais il *pleut* dehors. (*BLEU:* 68.1, *BLEURT:* 79.9, *COMET:* 0.97)
(2) I want to go get coffee but it's *closed* outside.	Je veux aller prendre un café mais il est *fermé* dehors. (*BLEU:* 70.7, *BLEURT:* 53.9, *COMET:* -0.15)
(3) I want to go get coffee but it's *pouring* outside.	Je veux aller chercher du café, mais ça textitcoule dehors. (*BLEU:* 51.9, *BLEURT:* 31.6, *COMET:* -0.38)
(4) I want to go get coffee but it's *been* outside.	Je veux aller prendre un café, mais ça a *été* dehors. (*BLEU:* 68.4, *BLEURT:* 27.1, *COMET:* -0.88)
(5) I want to go get coffee but it's *starting* outside	Je veux aller prendre un café, mais ça *commence* dehors. (*BLEU:* 68.5, *BLEURT:* 31.0, *COMET:* -0.57)
LM Infill + SSI interpretation & translation:	
(1) I want to go get coffee but it's *cold* outside.	Je veux aller prendre un café, mais il fait *froid* dehors. (*BLEU:* 90.3, *BLEURT:* 92.7, *COMET:* 1.20)
(2) I want to go get coffee but it's *warm* outside.	Je veux aller prendre un café mais il fait *chaud* dehors. (*BLEU:* 78.1, *BLEURT:* 79.1, *COMET:* 1.12)
(3) I want to go get coffee but it's *driving* outside.	Je veux aller prendre un café mais il *conduit* dehors. (*BLEU:* 70.4, *BLEURT:* 26.5, *COMET:* -0.69)
(4) I want to go get coffee but it's *closing* outside.	Je veux aller prendre un café mais il se *ferme* dehors. (*BLEU:* 69.8, *BLEURT:* 23.2, *COMET:* -0.81)
(5) I want to go get coffee but it's *dark* outside.	Je veux aller prendre un café, mais il fait *noir* dehors. (*BLEU:* 82.3, *BLEURT:* 73.7, *COMET:* 0.80)

Table 4.10: Continuation of Table 4.9. Examples of machine translation of slang.

4.7.2 Results

Figure 4.4 summarizes the results. Overall, the semantically informed approach tends to outperform the baseline approaches for the range of top retrievals (from 1 to 20) under all three metrics considered, with the exception of BLEURT evaluated on German where the semantically informed approach gives very similar performance as the language model baseline. While not all predicted interpretations correspond to the ground-truth definitions, the set of interpreted sentences often contain plausible interpretations that result in improved translation of slang. Table 4.7 to 4.10 provide some example translations. We observe that quality translations can be found reliably with a small number of interpretation retrievals (i.e., around 5) and the quality generally improves as we retrieve more candidate interpretations. However, it is still difficult to make a notable improvement with a single retrieval and machine translation of slang remains an open problem. Nevertheless, our approach may be integrated with a slang detector (e.g., Pei et al. 2019) to recommend high-quality translations in natural context that involves slang.

4.8 Conclusion

We have presented the first principled framework for automated slang interpretation that takes into account both contextual information and knowledge about semantic extensions of slang usage. We showed that our framework is more effective in interpreting and translating the meanings of English slang terms in natural sentences in comparison to existing approaches that rely more heavily on context to infer slang meaning. Future work could explore incorporating prior information into the interpretation framework. For example, the frequency of concepts expressed by slang, for that slang tends to express concepts such as drug and sex very frequently (Green, 2010; Eble, 2012).

Chapter 5

Semantic variation in slang

The contents of this chapter are based on my previous publication (Sun and Xu, 2022).

5.1 Motivation

Our computational framework thus far assumes that the same set of word-meaning associations of slang apply universally to all language users. However, the use of slang is often restricted to a specific group of users—a defining characteristic of slang that causes the meaning of a slang term to vary in different communities (Andersson and Trudgill, 1992; Mattiello, 2005; Eble, 2012). However, semantic variation in slang remains an under-explored topic in natural language processing and no formal model of the variation process has been proposed. This chapter explores semantic variation in slang by focusing on characterizing regularity in the geographical variation of slang usages attested in the US and the UK over the past two centuries. We show how the modeling of slang as sense extension can help explain the driving forces behind semantic variation in slang.

We define semantic variation in slang as the difference in meaning of a slang term across different communities. For example, Figure 5.1 shows an example where the commonly-used slang word *beast* has divergent meanings in different regions (or more specifically, two different countries in this case). Whereas it is more often used to

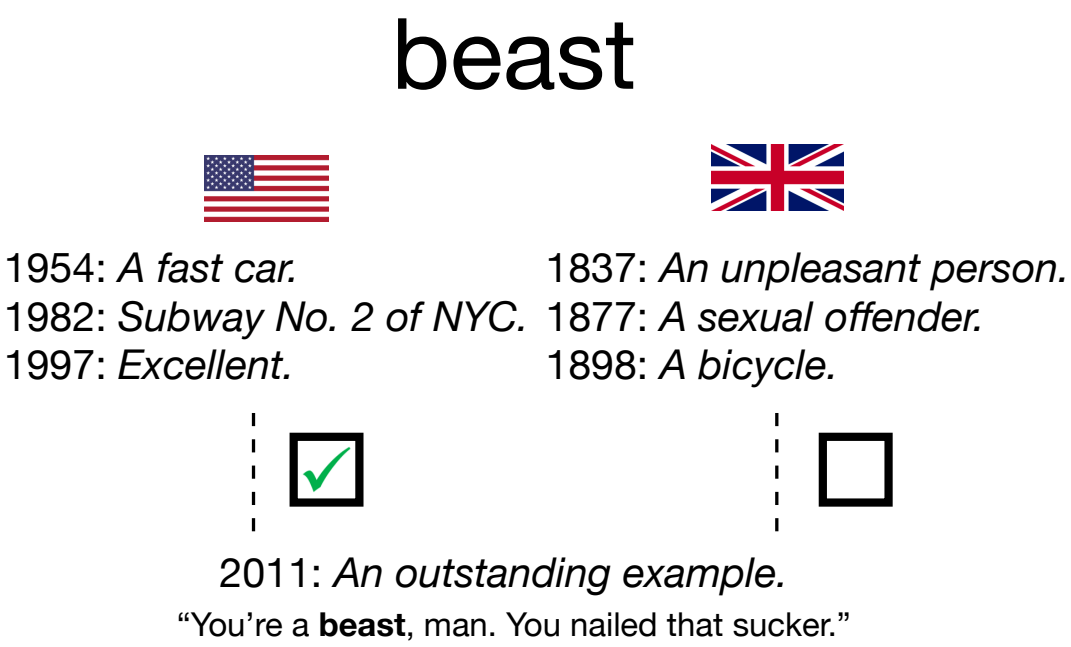


Figure 5.1: Illustration of semantic variation in the slang word *beast*, with senses recorded in American and British English respectively. We develop models of semantic variation in slang and evaluate them on a region-inference task: For a newly emerged slang sense, infer its regional identity given the slang term's historical usages from different regions.

express positive things or sentiment in the US, the same slang word has been used to express more negative senses in the UK.

Recent work has quantified semantic variation in non-standard language of online communities using word and sense embedding models and discovered that community characteristics (e.g., community size, network density) are relevant factors in predicting the strength of this variation (Del Tredici and Fernández, 2017; Lucy and Bamman, 2021). However, it is not clear *how* slang senses vary among different communities and what might be the driving forces behind this variation.

As an initial step to model semantic variation in slang, we focus on regional semantic variation between the US and the UK by considering a regional inference task illustrated in Figure 5.1: Given an emerging slang sense (e.g., 'An outstanding example') for a slang word (e.g., *beast*), infer which region (e.g., US vs. UK) it might have originated from based on its historical meanings and usages. Our premise is that

a model capturing the basic principles of semantic variation in slang should be able to trace or infer the regional identities of emerging slang meanings over time. Our experiments focus on the regional semantic variation between the US and the UK but the proposed models are widely applicable to the modeling of semantic variation in slang across more fine-grained communities.

5.2 Theoretical Hypotheses

We consider two theoretical hypotheses for characterizing regularity in semantic variation in slang: communicative need and semantic distinction. We evaluate these theories using slang sense entries from Green's Dictionary of Slang (GDoS, Green, 2010) over the past two centuries. Analysis on GDoS entries is appropriate because 1) a more diverse set of topics is covered compared to domain-specific slang found in online communities (e.g., Reddit), and 2) the region and time metadata associated with individual sense entries support a diachronic analysis on the semantic variation in slang. To preview our results, we show that both communicative need and semantic distinction are relevant factors in predicting semantic variation in slang, with an exemplar-based chaining model offering the most robust results overall. Meanwhile, the relative importance of the two factors is time-dependent and fluctuates over different periods of history.

5.2.1 Communicative need

Prior work has suggested that slang may be driven by culture-dependent *communicative need* (Sornig, 1981). We refer to communicative need as how frequently a meaning needs to be communicated or expressed within a given community. Following recent work (e.g., Kemp and Regier 2012; Ryskina et al. 2020), we estimate communicative need based on usage frequencies from Google Ngram[1] over the past

[1] https://books.google.com/ngrams

two centuries.[2] In the context of semantic variation in slang, certain things might be more frequently talked about in one region (or country) over another. As such, we might expect these differential needs to drive meaning differentiation in slang terms. For example, a US-specific slang sense for *beast* describes the subway line No. 2 of the New York City transit network. The need to communicate this specific subway line was presumably high in the US due to frequent crime cases on the train.[3] On the other hand, this sense would not be relevant to residents of the UK given its specificity to an US entity.

5.2.2 Semantic distinction

We also consider an alternative hypothesis termed *semantic distinction* motivated by the social functions of slang (cf. Labov, 1972; Hovy, 1990)—language that is used to show and reinforce group identity (Eble, 2012). Under this view, slang senses may develop independently in each region and form a semantically cohesive set of meanings that reflect the cultural identity of a region. As a result, emerging slang senses are more likely to be in close semantic proximity with historical slang senses from the same region.[4] For example, the slang *beast* has many senses in the US that describe something virtuous while senses in the UK often describe criminals. An emerging sense such as 'An outstanding example' would be considered more likely to originate from the US due to its similarity with the historical US senses of *beast*. Here we operationalize semantic distinction by models of semantic chaining from work on historical word meaning extension (Ramiro et al., 2018; Habibi et al., 2020), where each region develops a distinct chain of related regional senses over history.

[2]We acknowledge that experiment-based methods for estimating need exist (see Karjus et al., 2021), but these alternative methods are difficult to operationalize at scale and in naturalistic settings required for our analysis.

[3]https://www.barrypopik.com/index.php/new_york_city/entry/the_beast_2_subway_line

[4]It is worth nothing that communicative need and semantic distinction may not be completely orthogonal. In fact, differences in communicative need may drive semantic distinction. However, we consider these hypotheses as alternative ones because they are motivated by different functions.

Category	Tags
Slang	Cockney, informal, slang, vulgar
US	Boston, California, Florida, Louisiana, Maine, Midwestern-US, New-Jersey, New-York, New-York-City, North-America, Northern-US, Pennsylvania, Philadelphia, Southern-US, Texas, US, Virginia, in US, in US and Canada, in US usually formal, in the US
UK	Britain, British, Cornwall, Derbyshire, Devon, East-Anglia, England, Kent, Liverpudlian, Mackem, Midlands, Multicultural-London-English, Norfolk, Northern-England, Northern-English, Northumbria, Orkney, Oxford, Pembrokeshire, Scotland, Shetland, Teesside, Tyneside, UK, Ulster, Wales, West-Midlands, Yorkshire, in Britain, in UK, of England

Table 5.1: Wiktionary metadata tags used to determine whether a sense is a slang or belongs to US or UK.

5.3 Quantifying variation in slang

5.3.1 Slang vs. conventional

We first compare slang with conventional language by counting the number of sense entries in the English Wiktionary.[5] We extract all word entries using WikExtract (Ylonen, 2022) and only consider those that have 1) at least one slang sense and 2) senses in both US and UK. We determine whether a sense is regional and/or slang using metadata tags associated with each sense. Table 5.1 shows the full list of tags used. If one of the tags is found in the metadata, then the entry is considered part of the corresponding category. Entries with both US and UK tags are considered neither US-specific nor UK-specific. We obtain 810 slang words after filtering using the criteria above which contain 8,769 conventional senses and 1,262 slang senses. The proportion of senses with regional tags is shown in Figure 5.2, confirming that semantic variation is much more prevalent in slang compared to conventional language.

5.3.2 Regional slang

5.3.2.1 Data collection

We collect slang lexical entries from Green's Dictionary of Slang (GDoS, Green, 2010),[6] a historical English slang dictionary covering more than two centuries of slang

[5] https://en.wiktionary.org/
[6] https://greensdictofslang.com/

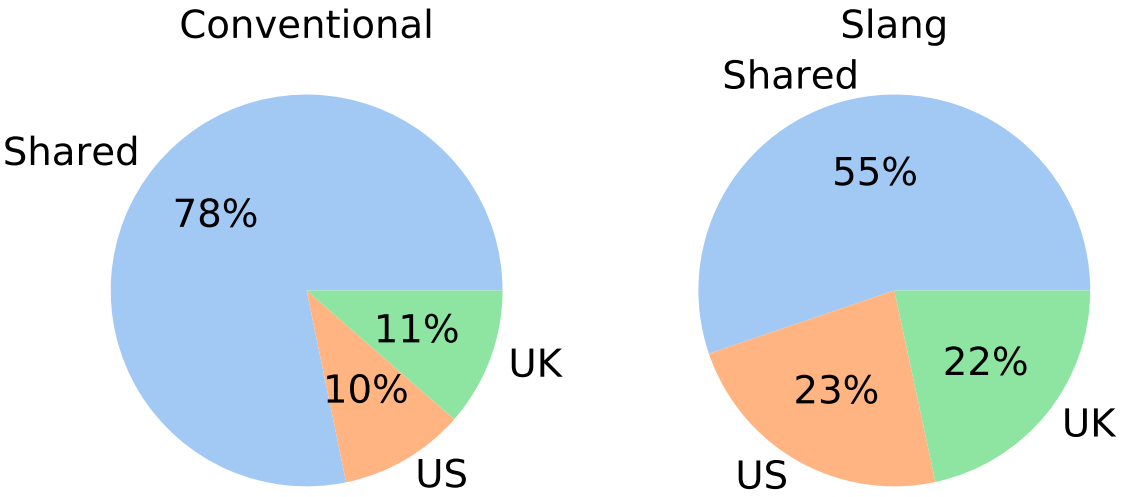


Figure 5.2: Distribution of regional identities among sense entries found in the English Wiktionary.

usage. Each word entry (e.g., *beast*) in GDoS is associated with one or more sense entries. We consider each sense entry as a data point in our analysis. A sense entry contains a definition sentence (e.g., 'An outstanding example.') and a series of references. Each reference contains a region tag (e.g., US or UK), a date tag (e.g., 2011), and a sentence indicating the origin of the reference. In some cases, the reference contains an example usage sentence of how the slang is used in context.[7]

We collect all sense entries with at least one valid reference. A reference is considered valid if both its region tag and date tag are not missing nor invalid. For each reference, we automatically extract the associated context sentence and consider one to be valid if it contains precisely one exact occurrence of the word in the sentence. If a valid context sentence is found then it is attached to the corresponding reference. The resulting sense entry may have none or more than one context sentences. In the latter case, we select the context sentence with the earliest time tag to be associated with the sense entry, so that it best represents the usage context of when the sense first emerges. The earliest time tag found in the references is considered the time of emergence for a sense entry. We filter all abbreviation entries as these entries don't create new meaning. In the case of a homonym (i.e., multiple word entries

[7]We choose GDoS over alternative resources (e.g., Lewin and Lewin, 1988; Dalzell and Partridge, 2009; Ayto and Simpson, 2010) because it covers a diverse set of slang usages from different regions and time periods.

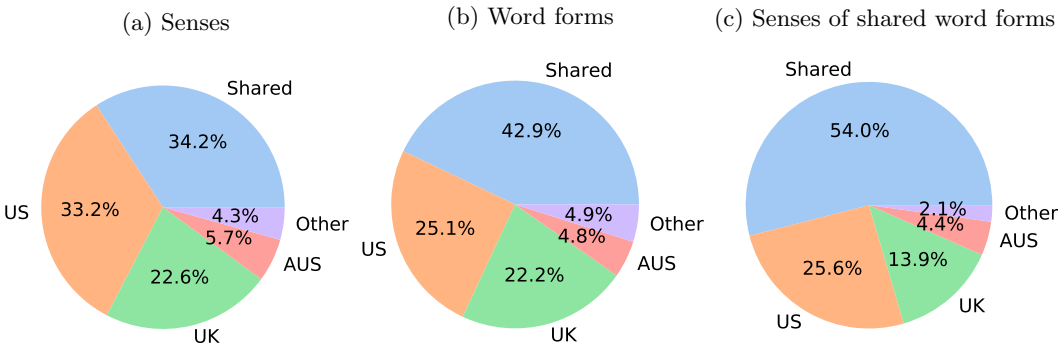


Figure 5.3: The distribution of GDoS slang senses and word forms across different regions. A word or sense is considered shared if two or more distinct region tags can be found in the constituent references.

for the same word form), we collapse all entries into a single word entry. After pre-processing, we obtain 42,758 distinct words with 76,650 associated sense entries. On average, each sense entry contains 4.48 tags of attested time and region. We provide our pre-processing script in our Github repository[8] to facilitate reproduciblility.

5.3.2.2 Data analysis

We first analyze entries collected from GDoS to quantify semantic variation. For each sense entry, we determine its regional identity using the region tags associated with each reference. Note that there may be more than one valid region tag associated with each sense entry. In such cases, we consider the sense entry to be a shared sense across all constituent region tags. Otherwise, the sense entry is considered regional. Likewise, a word entry is considered shared if two or more distinct region tags can be found among any of its sense's references.

Figures 5.3a and 5.3b show the distribution of region identities across all sense and word entries in GDoS. We observe substantial lexical variation within the data where more than half of the word forms are regional. While most of the sense entries are also regional, many of them may be associated with regional word forms. In this case, the variation is caused by difference in lexical choice and does not entail semantic

[8]Code and data scripts available at: https://github.com/zhewei-sun/slangsemvar

variation. We control for lexical variation by only considering sense entries associated with shared word forms. This results in 48,565 sense entries with an average of 5.80 tags per entry. Figure 5.3c shows the distribution of the resulting region identities. We observe that even after controlling for lexical variation, roughly half of the senses remain regional. Moreover, much of the semantic variation is captured by the US and UK regions, with Australian slang also making up a notable portion. We therefore focus on modeling semantic variation between the two most represented regions.

5.4 Models of semantic variation in slang

5.4.1 Predictive task

We model semantic variation by formulating a regional inference task: Given an emerging slang sense s for a word w, infer the region $r \in \mathcal{R}$ from which the emerging sense originates. Here, $\mathcal{R}$ is the set of regions being considered and an example would be the set {US, UK}. A semantic variation model $\mathcal{V}$ is then defined as follows:

$$P(r) \propto \mathcal{V}(s, w, r) \tag{5.1}$$

Here, the semantic variation model $\mathcal{V}$ captures the likelihood of observing the emerging slang sense s expressed using word w within region r and can be either generative or discriminative in nature. Given the semantic variation model, the target region can be predicted by maximizing the likelihood:

$$r^* = \arg\max_{r \in \mathcal{R}} \mathcal{V}(s, w, r) \tag{5.2}$$

An effective semantic variation model would prefer regions that are more likely for the new sense to emerge. We next describe models of semantic variation $\mathcal{V}$ motivated by both communicative need and semantic distinction.

5.4.2 Models based on communicative need

We first describe a set of semantic variation models $\mathcal{V}$ inspired by the communicative need principle. Under this hypothesis, language users in different communities need differing expressions to express concepts of particular interest to the community (Sornig, 1981). We operationalize communicative need using frequency statistics of historical corpora from each region. First, we propose a *form frequency* model that considers the frequency of the slang word form w:

$$\mathcal{V}_{\text{form\_frequency}}(s, w, r) \propto f(w; r, s_t - \alpha : s_t) \tag{5.3}$$

Here, $f(w; r, s_t - \alpha : s_t)$ is the frequency of observing word w from region r within a time window α strictly preceding the sense's time of emergence s_t. Note that the form frequency model does not take into account any semantic information from the emerging sense s and simply estimates whether the word form w is more prevalent in one region.

The *semantic need* model incorporates semantic information by checking the frequency of all content words within the definition sentence s_d of sense s:

$$\mathcal{V}_{\text{semantic\_need}}(s, w, r) \propto \sum_{c \in \text{content}(s_d)} f(c; r, s_t - \alpha : s_t) \tag{5.4}$$

Words in the sense definition sentence reflect concepts that are highly relevant to the slang's intended meaning. The semantic need model thus quantifies the need for these concepts with respect to each community.

The *context need* model is informed by the usage context sentence s_c of sense s:

$$\mathcal{V}_{\text{context\_need}}(s, w, r) \propto \sum_{c \in \text{content}(s_c) \setminus w} f(c; r, s_t - \alpha : s_t) \tag{5.5}$$

We remove the word w since it is not part of the context. The context need model checks the communicative context to estimate contextual relevance with respect to

each region. Specifically, it takes into account the concepts that co-occur with the slang usage and their respective need in each community.

Both of the above models can also be framed as a majority vote model instead of taking a sum of frequencies:

$$\mathcal{V}(s,w,r) \propto \sum_{c} \begin{cases} 1, & \text{if } \max_{r'} f(c;r',s_t-\alpha:s_t) = f(c;r,s_t-\alpha:s_t). \\ 0, & \text{otherwise.} \end{cases} \tag{5.6}$$

Here, c is the set of content words relevant to each model. For each content word, a "vote" is contributed to a community if the corresponding word frequency is the highest among all communities. We find the majority vote scheme to be robust in our experiments as frequency counts of common words could otherwise dominate the estimates.

5.4.3 Models based on semantic distinction

Slang semantics may also diverge due to its social function, where language users in a community wish to create distinct senses to express their social identity (Eble, 2012). As a result, slang senses from different communities might evolve into cohesive but distinct clusters. For example, many slang senses of *beast* are used to express positive concepts in the US but negative ones in the UK. Here, the cluster of senses are internally conhesive (i.e., many US senses describing positive concepts) but distinct across the two communities (i.e., positive vs. negative connotations for US vs. UK). Motivated by this hypothesis, we model semantic variation using historical slang senses associated with the word w in a region r that emerged before s_t, denoted as $\mathcal{S}_{w,s_t,r}$. Under this paradigm, the semantic variation model $\mathcal{V}$ can be specified as follows:

$$\mathcal{V}_{\text{distinction}}(s,w,r) \propto g(s,\mathcal{S}_{w,s_t,r}) \tag{5.7}$$

Here, the function g can be viewed as a classifier that measures the categorical similarity between the emerging sense s and historical senses from region r. We model g generatively using semantic chaining models from historical word sense extension which are motivated by mechanisms of human categorization (Rosch, 1975; Nosofsky, 1986).[9] We adapt three prominent variants of semantic chaining from Ramiro et al. (2018): 1) the *one nearest neighbor (onenn)* model that only considers the most similar historical sense; 2) the mean *exemplar* model that accounts for all historical senses; and 3) the *prototype* model which collapses all historical senses into a single prototypical sense. When performing chaining, each sense is represented by embedding its corresponding definition sentence s_d using a sentence embedder E:

$$g_{\text{onenn}}(s, w, r) = \max_{s' \in \mathcal{S}_{w,s_t,r}} sim(E(s_d), E(s'_d)) \tag{5.8}$$

$$g_{\text{exemplar}}(s, w, r) = \frac{1}{|\mathcal{S}_{w,s_t,r}|} \sum_{s' \in \mathcal{S}_{w,s_t,r}} sim(E(s_d), E(s'_d)) \tag{5.9}$$

$$g_{\text{prototype}}(s, w, r) = sim\left(E(s_d), \frac{1}{|\mathcal{S}_{w,s_t,r}|} \sum_{s' \in \mathcal{S}_{w,s_t,r}} E(s'_d)\right) \tag{5.10}$$

The similarity between two sense embeddings is computed using negative exponentiated distance with a learnable kernel width parameter h:

$$sim(e, e') = \exp\left(-\frac{||e - e'||_2^2}{h}\right) \tag{5.11}$$

When data is available, the kernel width parameter h can be optimized by constructing training examples from the full set of historical senses $\mathcal{S}_{w,s_t}$

[9] Sentential context can be potentially integrated to achieve higher accuracies but here we focus on senses alone to examine the effect of semantic cohesiveness operationalized by cognitively motivated modeling approaches.

k	Word entries	US senses	UK senses	Shared senses	Test set
3	388	3273	1889	2007	1722
4	209	2063	1263	1272	1200
5	114	1342	827	877	788
6	64	842	550	577	548
7	44	627	423	446	424
8	30	455	316	337	310
9	21	286	239	240	230
10	14	192	176	156	162

Table 5.2: Number of GDoS word and sense entries obtained after constraining the minimum number of regional senses per region (k). Senses are divided into regional and shared based on region tags associated with sense references. The last column shows the sizes of the test sets where each is composed of an equal number of test senses from each region.

5.5 Experiments

5.5.1 Setup

We test our semantic variation models on region inference using GDoS word entries that show high regional variation in their senses. Specifically, we consider all word entries with at least k regional senses that have emerged after 1800 in each region of interest. We consider $k \in [3, 10]$ and Table 5.2 shows the number of words and senses that match the criteria for each k when considering {US, UK} as the set of regions.

All senses emerged after 1900 are treated as a time series of test examples. For example, all senses of a word that emerged before 1900 will be used as historical senses when predicting the region for the first sense post 1900 and this sense will then be considered as a historical sense when making a prediction for the subsequent sense. Word entries with sparse regional senses (i.e., $k = 1$ or $k = 2$) are excluded because they often result in uninformative test examples where not a single slang usage in one region is available prior to 1900.

Since there are often a disproportionate number of test senses between the two regions, we create class-balanced test samples by subsampling eligible test senses in each time series. For example, a word entry with 5 US sense entries and 3 UK sense entries emerged after 1900 will result in 6 test examples where 3 out of the 5 US

senses are randomly sampled while all UK senses are kept. Even if a sense entry has not been sampled for prediction, it will still appear in the history when predicting subsequently emerged senses. The last column of Table 5.2 shows the sizes of the class-balanced test samples where half of the sense entries come from each region. To account for all senses in the data, we repeat the sampling procedure 20 times in all of our experiments and report the mean predictive accuracy. Word lists for each k can be found in our Github repository.

We use case-insensitive normalized frequency from the 2019 version of Google Ngram's "American English" and "British English" corpora to estimate word frequencies for all communicative need models and set the window size α to 10 years. The list of stopwords from NLTK (Bird et al., 2009) is used to filter for content words. We apply additive smoothing of $1e^{-8}$ and 1 to normalized frequency and majority vote models respectively. In the case of a tie, the model defaults to predicting US.

For semantic distinction based models, sense embeddings are obtained by embedding their respective definition sentences from GDoS using Sentence-BERT (SBERT, Reimers and Gurevych, 2019). In addition to the semantic chaining models, we include *LDA* and *logistic regression* as discriminative baselines for the classifier g in Equation 5.7 where each sense's definition sentence is encoded using SBERT and used as the feature vector. We also include a *sense frequency* baseline that always predicts the most frequent sense tag observed in the historical senses. When all historical senses correspond to a single region label, then that label is taken as the prediction for all semantic distinction models.

To train the kernel width parameter in semantic chaining (i.e., h in Equation 5.11), we consider all historical senses as a time series and train on as many predictions as data allows. For example, if a list of senses emerged prior to the to-be-predicted sense, then we iterate through these senses in their order of emergence. As soon as there is an observation from each class, chaining probabilities are estimated and the negative log-likelihood of the corresponding region is included in the loss function for

	No shared senses			With shared senses		
Model	US senses	UK senses	All senses	US senses	UK senses	All senses
Form frequency	49.9 (1.20)	50.7 (0.20)	50.3 (0.57)			
Semantic need	54.0 (1.77)	50.1 (0.43)	52.1 (0.91)		Not applicable	
Context need	65.9 (1.37)	43.8 (0.46)	**54.8** (0.67)			
Sense frequency	55.9 (1.44)	34.1 (0.26)	45.0 (0.75)	59.2 (1.33)	34.0 (0.27)	46.6 (0.66)
LDA	51.7 (1.37)	45.3 (0.35)	48.5 (0.73)	54.7 (1.46)	45.5 (0.50)	50.1 (0.81)
Logistic reg.	52.5 (1.57)	40.0 (0.35)	46.2 (0.84)	56.5 (1.24)	39.3 (0.32)	47.9 (0.64)
Onenn	60.9 (1.35)	53.0 (0.42)	56.9 (0.71)	72.4 (1.32)	38.0 (0.41)	55.2 (0.64)
Exemplar	60.1 (1.31)	57.8 (0.40)	**58.9** (0.70)	60.0 (1.66)	58.6 (0.38)	**59.3** (0.85)
Prototype	57.6 (1.38)	53.1 (0.37)	55.4 (0.77)	60.3 (1.71)	54.7 (0.30)	57.5 (0.88)

Table 5.3: Mean percentage accuracy of all models on the region tracing task for post 1900 senses associated with words that have at least 5 regional senses in each region (US and UK; $k = 5$). Standard deviation of accuracies taken across 20 test samples is shown in parenthesis. The right-hand side shows the results after including shared senses in both training and prediction.

optimization. We use L-BFGS-B (Byrd et al., 1995) to optimize the kernel width with a default value of 1 and a bound in [0.01, 100].

5.5.2 Inferring regional identity of slang

We now evaluate both the communicative need and semantic distinction based semantic variation models on the regional inference task. Table 5.3 shows the mean predictive accuracy of all models on the $k = 5$ test set. For communicative need, we observe that both the semantic need and context need models made better predictions than the simple form frequency model and the random baseline (i.e., 50% accuracy). For semantic distinction, the chaining models consistently outperform both the random baseline and the discriminative classifiers.

Since only a few historical senses are available for training, the standard classifiers' (LDA and logistic regression) suffers from data sparsity which result below-baseline performance. Meanwhile, both models of slang variation are able to leverage enriched data points from history to perform well in a few-shot setting. Indicated by poor performance from the sense frequency baseline, we observe no discernible patterns in the emergence trajectory of senses across regions when the content of the slang usage is disregarded. In line with previous applications of semantic chaining to linguistic

categories (Habibi et al., 2020; Yu and Xu, 2021), we also find exemplar-based chaining performing the best among alternative chaining models, suggesting that regional slang senses tend to form cohesive neighborhoods in the underlying semantic space. Both the exemplar and prototype models also tend to rely less on sense frequency and produce more balanced predictive accuracies across the two regions.

We also consider the inclusion of shared senses in addition to their regional counterparts. As shown in Table 5.2, shared senses account for a large portion of the sense inventory that could result in more data. For each shared sense, we track its list of references to determine its regional identity at a particular point in history. For example, a sense containing a US reference in 1930 and a UK reference in 1940 would be considered US exclusive when used to predict the region of a sense emerging between 1930 and 1940. Senses with shared regional identities (e.g., the aforementioned sense after 1940) are considered by both regional categories in semantic chaining models to obtain more accurate kernel width estimates. We observe better model performance in all models that consider historical senses after including shared senses. Introducing shared senses most notably improved the prototype model where more senses arguably led to more accurate estimates of the prototypical senses.

Figure 5.4 shows the predictive accuracy of the best performing models over all samples of k. Overall, both communicative need and semantic distinction models are able to capture a notable amount of variation in the data, with the semantic chaining based models giving the best predictability. Also, the advantage of chaining-based models over frequency-based models diminishes for more polysemous slang (which presumably is also more frequently used). This suggests that as a slang obtains more senses, its set of senses becomes less cohesive and the slang word is more likely to be used to express concepts with high communicative need that are more coarsely related to its historical meaning. An alternative explanation is that those historical senses become conventionalized or dismissed over history and are thus no longer relevant in the emergence of new slang senses. Next, we test this hypothesis by constraining the

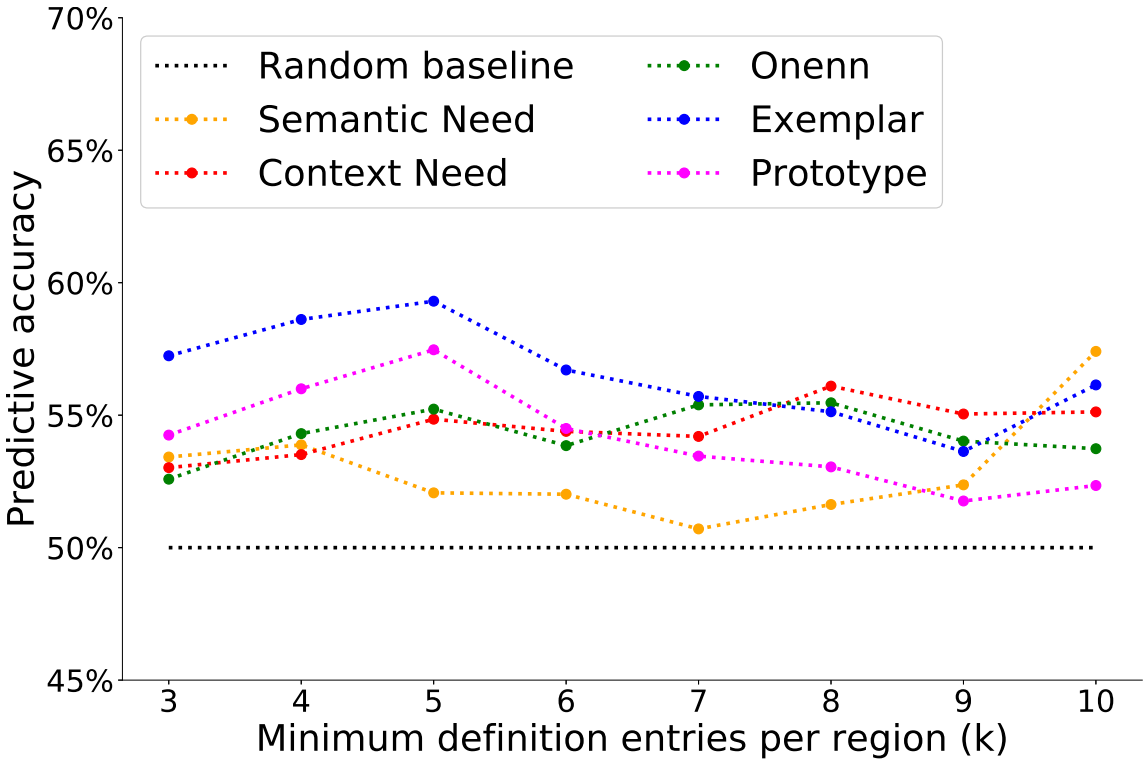


Figure 5.4: Predictive accuracy of the best performing models relative to the minimum number of regional senses (k) in sampled word entries. All shared historical senses are used in semantic chaining.

Model	US senses	UK senses	AUS senses	All senses
[No shared senses]				
Sense frequency	58.4 (2.58)	25.3 (1.95)	4.8 (0.78)	29.5 (1.01)
LDA	58.1 (3.56)	23.4 (1.42)	11.8 (1.29)	31.1 (1.38)
Logistic reg.	54.3 (3.33)	25.5 (1.74)	10.1 (1.08)	30.0 (1.34)
Onenn	52.4 (2.80)	26.2 (1.75)	28.3 (1.52)	35.6 (1.15)
Exemplar	42.1 (3.22)	30.7 (1.52)	38.9 (1.68)	**37.2** (1.34)
Prototype	50.4 (3.26)	29.4 (1.71)	22.4 (1.64)	34.0 (1.34)
[With shared senses]				
Sense frequency	54.3 (2.85)	31.3 (1.58)	1.2 (0.42)	28.9 (1.11)
LDA	40.5 (3.28)	23.4 (1.42)	11.8 (1.29)	25.3 (1.21)
Logistic reg.	56.5 (2.97)	26.5 (1.74)	7.8 (1.11)	30.2 (1.17)
Onenn	67.9 (2.74)	23.1 (2.17)	23.0 (1.03)	38.0 (1.21)
Exemplar	45.1 (3.06)	27.0 (1.49)	49.3 (1.70)	**40.5** (1.32)
Prototype	51.7 (3.50)	32.1 (1.95)	31.7 (1.20)	38.5 (1.56)

Table 5.4: Mean percentage accuracy of all models on the region tracing task for post 1900 senses associated with words that have at least 3 regional senses in each region (US, UK and AUS; $k = 3$). Standard deviation of accuracies taken across 20 test samples is shown in parenthesis.

number of senses considered in the chaining models.

Finally, we consider a 3-way classification experiment involving Australian slang,

which makes up a small but notable portion of GDoS. Here, we only consider words with at least 3 regional senses (i.e., $k = 3$) due to data sparsity. Also, communicative need models are not evaluated because frequency statistics are not available for the Australian region on Google Ngram. We find 44 word entries that match the criteria with 395, 254, and 167 regional entries for US, UK, and Australia respectively. We also include 467 shared senses similar to the experiment described in Section 5.5.2. We sample class-balanced test sets and obtain 309 examples evenly divided among the 3 regions. We repeat this sampling procedure 20 times to account for all senses. Table 5.4 shows the results. We observe similar trends as in the US and UK only case where the semantic chaining models substantially outperform all baselines. The exemplar model achieves the highest predictive accuracy both overall and on the Australian test cases despite the set of Australian historical senses being less frequent than the others.

5.5.3 Memory in semantic variation

Slang senses are known to be short-lived and become conventionalized or dismissed over time (Eble, 1989). We measure to what extent historical senses are relevant in the process of variation. We do so by constraining the number of historical senses seen by the chaining models based on their year of emergence. We focus on the $k = 5$ case and find that without a memory constraint, the average age of historical senses ranges from 36.8 years for test senses in the 1910s to 73.7 years for those in the 2010s. Figure 5.5 shows the mean predictive accuracy for all chaining models after removing historical senses that exceed the memory threshold. To preserve model efficacy, historical senses can still be used as examples to train the kernel width parameter, but those examples themselves are also restricted to historical senses within the memory threshold when making predictions during training. Despite our intuition, we observe a consistent upward trend in predictive accuracy as the memory constraint becomes more relaxed. Historical slang senses dating over 100 years nevertheless remain rele-

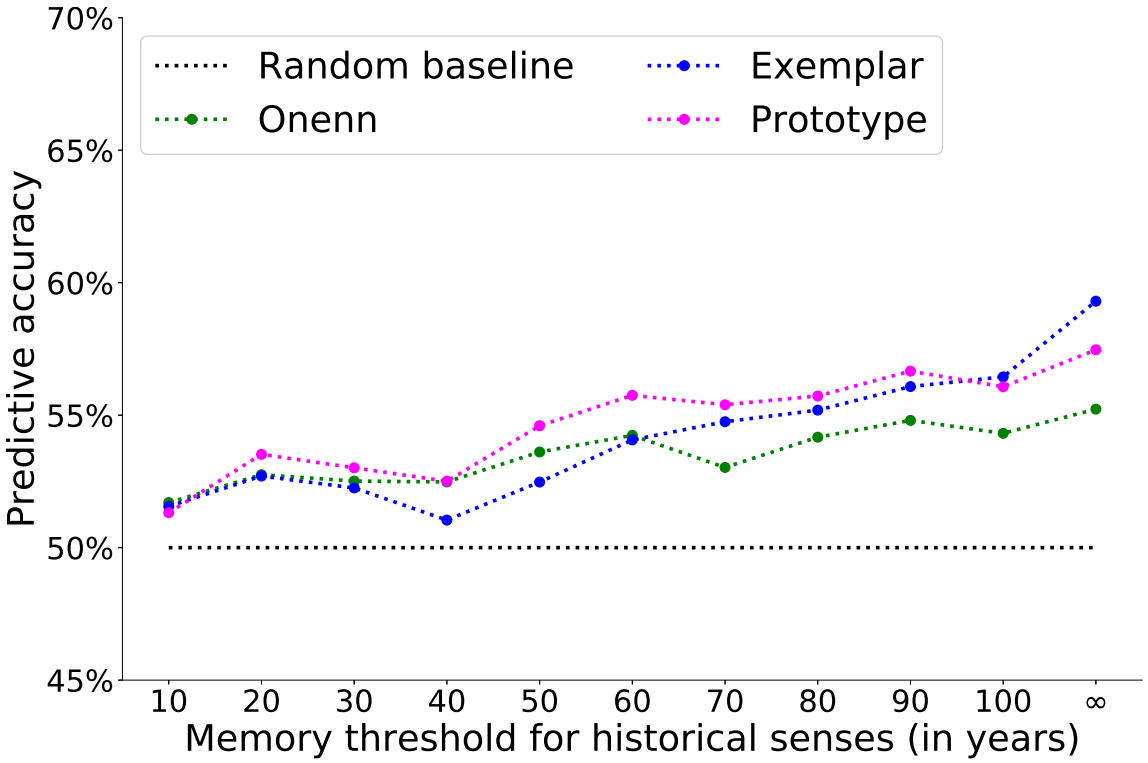


Figure 5.5: Predictive accuracy of all chaining models with shared senses after removing historical senses that exceed the memory threshold during prediction.

vant when considering the semantic variation of contemporary slang.

5.6 Conclusion

We have presented a principled and large-scale computational analysis of semantic variation in slang over history. Inspired by the theoretical hypotheses of communicative need and semantic distinction, we develop a computational approach to test these theories against regional slang attested in the UK and the US over a long period of history. We find regular patterns in semantic variation in slang that are predictive of regional identities of the emerging slang senses. While both hypotheses are found to be relevant for predicting semantic variation of slang, we observe that semantic distinction better explains the semantic variation in slang terms used in the US and the UK over the past two centuries. Our work sheds light on the basic principles of semantic variation of slang and provides opportunities for incorporating histori-

cal cultural elements in the automated processing of informal language such as slang generation and interpretation. Future work could consider extending the experiments toward more fine-grained community structures. Examples include the examination of slang usages across different US States (cf. Cassidy, 1985) or online communities of practice (cf. Del Tredici and Fernández, 2017, 2018).

Chapter 6

Conclusion

6.1 Summary

This dissertation explored the modeling of slang semantics and its use in improving natural language processing for slang. While existing approaches disregard regularities in slang usages or are not built specifically for slang, my contributions have focused on theoretically grounded approaches that leverage principles of semantic extension. My results show that the semantic extension of slang is indeed non-arbitrary, motivating a principled computational framework of slang semantics that addresses many challenging aspects of slang in NLP.

Chapter 1 introduced the essential tasks in NLP for slang and how characteristics of slang make these problems difficult. In this chapter, I also surveyed relevant linguistic work that not only characterizes what slang is but also lays out important challenges that must be addressed in achieving effective processing of slang. Chapter 2 reviewed existing NLP work on the automatic processing of slang. Although much of the existing work does not focus specifically on slang, but it remains relevant in studying its variation in online social media as well as the modeling of slang word formation. Finally, I also surveyed relevant cognitive and linguistics work on word sense extension that motivates the models I present in this dissertation.

The next three chapters described my core contributions to improve NLP for slang.

Motivated by by a combination of probabilistic models and deep learning techniques, Chapter 3 presented my work on slang generation in which a principled framework of slang semantics captured patterns of semantic extension attested in historical slang reuse — the case where an existing word in the lexicon is chosen to express a new meaning in slang. The generative nature of the framework allows generalization towards novel slang. Also, contrastively learned sense representations better captures regularities in slang semantics compared to off-the-shelf embeddings trained on conventional language, closing the gaps illustrated in Figure 3.2.

In Chapter 4, I showed how this slang generation model can be applied to the more practical task of slang interpretation and translation. By incorporating semantic information of the slang expression into a general purpose interpretation/translation model, the systems can make more informed choices that address the inherent ambiguities of a context-based model. This approach is particularly advantageous because only a relatively small sample of attested slang usage is required to train the model, whereas existing end-to-end neural network based approaches require much larger slang datasets to achieve adequate performance despite the scarcity of data.

Chapter 5 of this dissertation modeled the processes underlying slang semantic variation. Motivated by linguistic theory (Sornig, 1981; Mattiello, 2005; Eble, 2012), I considered two hypotheses based on communicative need and semantic distinction. In a large-scale experiment on regional semantic variation of slang over US and UK, my results have shown that both hypotheses play a role in predicting the regional variation and that semantics alone reveals non-trivial predictability of a slang's regional identity. The results show that contextual information indeed plays an important role in shaping how a slang is used and opens promising avenues of future work in incorporating contextual information into NLP systems for slang.

6.2 Future extensions

6.2.1 Extending the model of slang semantics

A limitation of the semantic framework proposed in this dissertation is that it only models cases of slang reuse but not coinage (i.e. slang usage with novel word forms). Because of this, little semantic information can be leveraged in tasks such as slang interpretation for newly coined slang. While many of the existing methods are proposed or can be potentially applied to slang coinage (see Section 2.3 for a detailed review), the relevant modeling approaches that have been applied to slang remain rudimentary. For instance, Ni and Wang (2017) only used a character-level LSTM encoder to represent the semantics of a slang word before combining it with contextual information. Here, it would be interesting to explore the applicability of more modern NLP techniques (e.g., Sennrich et al. (2016), Pinter et al. (2017), and Kudo and Richardson (2018)) in processing out-of-vocabulary words (OOVs) and to evaluate the extent to which recent LLM based methods could capture meanings of newly coined slang.

A key question yet to be addressed is whether the semantic model of newly coined slang words is similar to that of their conventional counterparts. In the case of slang reuse, we have theoretical evidence showing that slang semantic extension is mechanistically similar to conventional semantic extension but differs in the distribution of extension devices. Similarly, it is conceivable that how the meanings of a newly coined slang is inferred from its constituent morphological units can substantially differ from conventional coinage. For example, conventional and slang blends may share different processes in extending senses of their constituents to create a blended sense. Pinter et al. (2020) show preliminary evidence where differences in BERT representations of lexical blends and their constituents are farther apart than those of compounds. Here, many of the lexical blends may be slang as it is a common device in slang word formation (Eble, 2012). Existing word formation and OOV processing

methods, however, rely on training examples reflecting existing conventional words. Because of this, these methods implicitly assume that slang word formation shares the same underlying semantic model as conventional word formation. Future work could explore whether such an assumption can be warranted, and if not, modeling techniques that capture the semantics of newly coined slang.

Extending this further, the use of slang can also appear in idiomatic phrases. In this dissertation, I have considered phrasal slang with fixed forms. For example, the slang phrase *night owl* has been considered as a single lexical item in my proposed models. As long as dictionary entries exist for the phrase, the semantic model does not need to distinguish between words and phrases. Slang usages, however, can also appear as creative compositions of words. For example, instead of the canonical phrase *night owl*, one may come up with a sentence such as "Danny is an owl at night". Here, the idiomatic phrase *owl at night* constitutes a slang usage similar to *night owl*. It would thus be interesting to explore modeling approaches for idiomatic expressions (e.g., Fazly et al., 2009; Liu and Hwa, 2018; Zeng and Bhat, 2021, 2022) to extend the semantic model for slang.

Another important avenue of extension would be to consider slang in a multilingual setting. Work presented in this dissertation has focused on the modeling and processing of English slang but slang is also widely used in other languages and cultures. The processing of slang in other languages presents a great challenge in modeling efficiency as data becomes more scarce compared to English. A potential solution to alleviate this challenge is to transfer the semantic knowledge learned from English slang to the other language. Recent multilingual word alignment methods (Artetxe et al., 2018; Jalili Sabet et al., 2020; Shi et al., 2021) operate on structural similarities in the underlying embedding space instead of relying on supervised word pairs. This allows direct transfer of knowledge into another language from a learned embedding space such as the contrastive sense embeddings presented in Chapter 3.

It is important to note that such an approach assumes that both the creation and

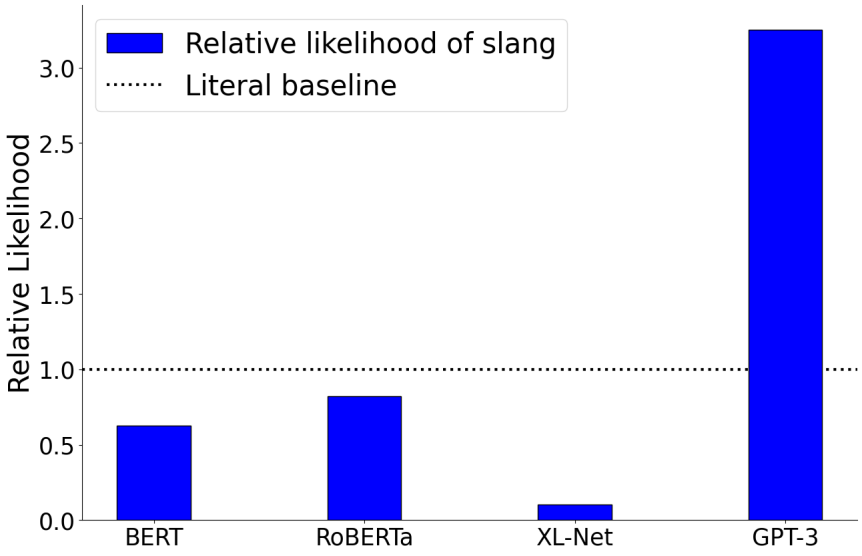


Figure 6.1: Relative language modeling performance between literal and slang usages for state-of-the-art large language models.

use of slang are similar across different language and cultures. While the methodology presented in this dissertation does not make explicit assumptions specific to a particular language, it is conceivable that differences in culture and values reflected in languages can lead to differences in how slang is created and used. For example, languages used by more progressive cultural groups may make more frequent use of innovative sense extension strategies. A careful evaluation on the results of transfer learning could reveal the validity of this assumption.

6.2.2 Slang and large language models

Recent advances in building large language models (LLMs) as foundational models for NLP have made significant advances in many important NLP tasks (Bommasani et al., 2021). Results from a set of preliminary experiments shown in Figure 6.1 and 6.2 show that this is also the case for processing slang. For Figure 6.1, I use 102 test sentences from the slang translation experiment in Chapter 4 (See Section 4.7 for details) where each sentence contains a marked slang and a corresponding literal paraphrase. The slang expression is then masked out and each language model is

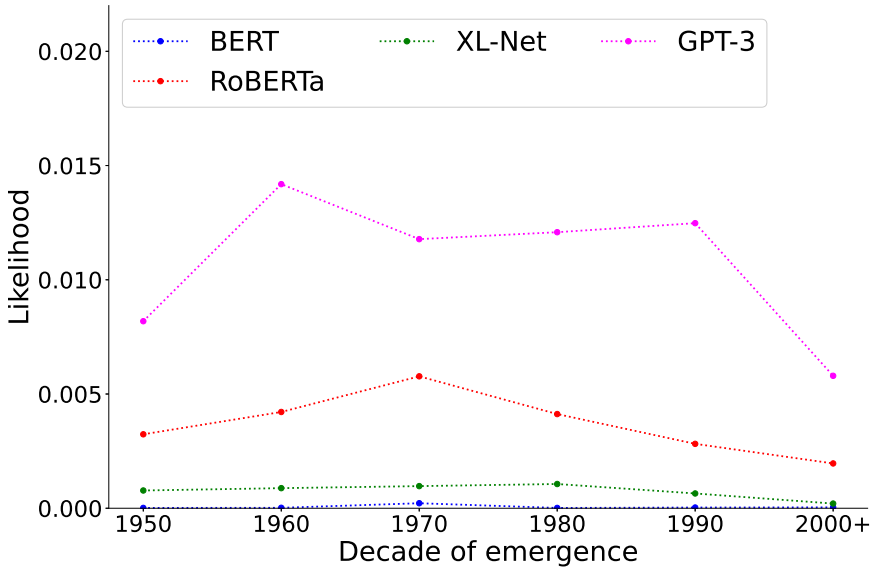


Figure 6.2: Absolute language modeling performance on slang usages for state-of-the-art large language models across different emergence period for the slang.

asked to fill in the blank. For example, LM probabilities of the slang word *blazing* and literal paraphrase *excellent* will be compared against:

1. Good purchase, that jacket is **blazing**.

2. Good purchase, that jacket is **excellent**.

Finally, the mean LM likelihood of all ground-truth literal and slang expressions are compared against:

$$\text{Relative likelihood} = \frac{\sum_i \mathcal{S}_i}{\sum_i \mathcal{L}_i} \tag{6.1}$$

Here, $\mathcal{S}_i$ denotes the language model probability assigned to the slang word in the i'th sentence and similarly $\mathcal{L}_i$ for the literal word's probability. Here, I aggregate over probabilities for each type instead of individual ratios to avoid over-emphasizing outlier slang that the model is either very confident or very impoverished on. To control for potential confounds in tokenization, I only consider sentences such that both the corresponding slang and literal paraphrase are single token expressions. For Figure 6.2, a set of 5,052 slang-containing sentences from Green's Dictionary of Slang

Model	OSD	GDoS	UD
fastText	0.35 ± 0.033	0.30 ± 0.010	0.31 ± 0.037
SBERT	0.32 ± 0.033	0.32 ± 0.010	0.28 ± 0.034
GPT-3	0.31 ± 0.032	0.31 ± 0.011	0.30 ± 0.035

Table 6.1: Normalized ranks (between 0 and 1, lower is better) of a word's slang definition embedding towards its conventional definition embedding over entries in The Online Slang Dictionary (OSD), Green's Dictionary of Slang (GDoS) and Urban Dictionary (UD). I compare the embeddings produced by GPT-3 against those computed in Chapter 3 using fastText (Bojanowski et al., 2017) and Sentence-BERT (SBERT; Reimers and Gurevych, 2019).

has been used. For each sentence, the LM probability of the slang expression is measured.

Although GPT-3 (Brown et al., 2020) is architecturally similar to the earlier BERT-like models (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019), access to much larger training data made it more confident on the slang usages compared to the literal paraphrases. An interesting direction of research is thus to interpret such large language models and find out whether models like GPT-3 are indeed modeling the semantics of slang. I repeat the sense representation experiment presented in Section 3.7.5 on GPT-3 sense representations.[1] Table 6.1 shows an extension of the results presented in Table 3.5. Although GPT-3 assigns much higher probabilities to slang terms, I do not observe any significant difference in the underlying geometry of the representations. The results suggests that GPT-3's source of knowledge comes from frequent instances of slang usage seen during training and simply treats them as additional "conventional" senses. It has yet been able to (or decided not to) encode any structural knowledge of slang into its representations.

The preliminary results shown may reflect memorization from data. Indeed, the slang usages being tested here have been well documented on the internet and have seen many years of usage. Higher likelihood on the slang expressions indicates that the model has higher confidence on the slang expression than the literal paraphrase. This suggests that GPT-3 can pinpoint the meaning and usage context of the slang quite well, something that is quite difficult to achieve without seeing the slang beforehand.

[1] Sense representations are obtained be encoding definition sentences using the *text-similarity-davinci-001* embedding model from the OpenAI API

Another piece of evidence suggesting data memorization is that GPT-3's performance on more contemporary slang is much higher than the competing models while it does not show such a high performance gap in historical slang. Once GPT-3 scale training becomes more accessible to researchers in the future, one can investigate the effect to which certain data sources have on slang related performance. For example, ablating parts of Urban Dictionary and Reddit data from training would allow us to pinpoint cases of data memorization.

6.2.3 Fairness and privacy

The use of slang is often associated with derogatory topics that can cause harm (Green, 2010; Eble, 2012). Previous study by Kulkarni and Wang (2017) has shown that slang usages not only reflect gender and religious stereotypes, they are often more socially prejudicial compared to conventional language. While many of these slang usages are explicit and thus easy to detect, there are many negatively-connotated usages that hide under the guise of a neutral word (e.g., using the word *published* to express 'Very ugly'). Recent approaches in detecting euphemistically coded language are often limited in scope, where the system can only perform detection and interpretation within a pre-determined set of topics (Magu and Luo, 2018; Zhu et al., 2021), a limitation caused by the lack of proper semantic representation beyond distributed semantic models. Naturally, the slang semantic model proposed in this dissertation can be applied to better decipher euphemistic uses of hate-speech.

It is also important to consider the potential biases and harm a system may introduce as a result of improved performance on slang. Language use reflects one's social identity (Clark, 1998; Eckert, 2012) and is especially salient in the case of slang (Labov, 1972, 2006; Mattiello, 2005; Eble, 2012; Slotta, 2016; Denis, 2021). For example, users belonging to certain social groups may be more likely to invoke a specific set of slang compared to the general public. Several issues may arise from such variations in language use. First, an NLP model that performs poorly on slang

compared to conventional language would introduce unwanted biases against social groups who are more likely to use slang, resulting in worse performance when used by these groups of users (cf. Blodgett and O'Connor, 2017). This issue can be alleviated by improving NLP models' abilities in processing slang to which this dissertation discusses potential methods to do so.

A more critical problem to ask is whether the NLP systems are biased towards specific sets of slang that correlate with users identities in certain social groups. Figure 6.2, for example, shows how LLMs such as GPT-3 may perform poorly on contemporary slang compared to more historical ones, resulting in performance bias against younger users who use contemporary slang more frequently. Note that this issue differs from the slang vs. conventional case. Even when the model can process slang in relatively similar performance compared to conventional language, certain sets of slang may be more favored than others. This could result from design decisions such as data selection. More recent slang usages, for example, may have less representation in a training corpus and therefore more difficult to process. It is important for future work to consider a framework that carefully evaluates potential biases relevant to slang usage and guide NLP practitioners towards design decisions that minimize such unwanted biases.

Aside from performance biases a model may introduce, user privacy is another important issue to consider. Since the use of slang may potentially reveal the social identity of the user, as shown in Chapter 5, it would be ideal to create an evaluation framework that determines the extent to which existing and future NLP systems can infer user identities. In cases where the models can reliably infer one's identity based on slang use, it may be of interest to introduce some form of differential privacy (Dwork and Roth, 2014) that prevents the leak of user identity. Here, instead of protecting a specific piece of information (e.g., a phone number), we would like to protect a user's identity regardless of the type of slang being used. For example, an embedding model could produce similar representations for the US slang *gucci*

and UK slang *massive* for expressing the meaning 'Good, excellent', making it more difficult for an LLM based system to infer the regional identity of the user. In this case, the ideal embeddings would capture the semantic similarities between *gucci* and *massive* but not to encode excessive demographic information associated with each slang usage.

6.2.4 Applications in linguistics and social science

While it is important to improve the performance of NLP systems for slang, the linguistic knowledge about slang that can be uncovered while building such systems is also worth careful attention. Most existing linguistic work on slang is qualitative in nature (e.g., Mattiello, 2005) or studies slang at a small scale (e.g., Denis, 2021). Although some of these studies also present data-driven results, the empirical evidence the conclusions were made from often only involves a handful of carefully picked examples. Warren (1992), being a rare exception, shows how large-scale data collection and processing can result in much more rigorous linguistic results. The collection of large-scale data used to train NLP systems for slang would allow linguists to conduct studies that analyze slang more diversely.

Successful NLP systems for slang can also reveal important characteristics of slang that has not been discussed in prior work. For example, my work on slang detection (Pei et al., 2019) has discovered that slang usages employ much more surprising Part-of-Speech shifts compared to conventional language. Also, the results of my slang generation work (Chapter 3) reinforces the findings that slang sense extension indeed shares similar underlying linguistic mechanisms with conventional sense extension, evident by above-chance performance from the chaining models using off-the-shelf embeddings. Furthermore, the substantial gain from employing contrastive learning suggests that the specific linguistic devices employed for sense extension indeed differs between slang and conventional language.

In addition to linguistic knowledge discovered in the process of enhancing NLP for

slang, the advancement in NLP for slang can also be directly applied to verify previously proposed linguistic theories using large-scale datasets. For example, Chapter 5 shows how a model of slang variation, combined with a large slang dictionary, can be applied to study slang variation at different periods of history. It shows how future work can apply slang NLP to uncover interesting linguistic insights at a scale that is not feasible with manual inspection.

Slang is ubiquitous in online social media text and thus presents research opportunities in computational social science. First, the ability to model slang may already bring noticeable improvement in related analytic tasks. Previous work has shown promising results in incorporating NLP methodologies for slang into sentiment analysis (Wu et al., 2018; Aly and van der Haar, 2020; Wilson et al., 2020), where the slang usages often reflect non-neutral sentiment. In Chapter 5, I also demonstrated how models of slang extension can be applied to model regional semantic variation of slang. At the same time, the use of slang is a good indicator of one's social identity (Labov, 1972, 2006; Mattiello, 2005; Eble, 2012; Slotta, 2016; Denis, 2021), allowing the inference of user identity in situations where such information is scarce, unreliable, and/or missing in our data. For instance, Chapter 5 of this dissertation demonstrates how one can trace the regional identity of a novel slang usage by looking at historical slang senses of the same slang expression. Such methodology can be leveraged to infer the regional identity of Reddit users based on their use of slang, the type of data often desired in online social media analysis but difficult to collect in practice.

Analysis on social media data also has good potential in further improving NLP for slang. Existing work has shown how contextual information alone (i.e. without text) can be used to predict user behavior (Waller and Anderson, 2020). Given that slang is a contextually motivated form of language, such contextual information can be incorporated to further enhance the performance of NLP systems for slang. The work presented in Chapter 5, for example, can inform the system about the slang user's regional identity, a piece of information that our system can leverage to output

more targeted results. For instance, if a user writes many US exclusive slang, then the subsequent user of the word *blazing* by the same user can be predicted to be expressing something good, the typical use of *blazing* in the US. Future work could extend this idea further by exploring different types of meta-information that are available in existing social media analyses (e.g., Gilbert and Karahalios, 2009; Del Tredici and Fernández, 2017; Waller and Anderson, 2020) and consider creating datasets of slang that capture such information.

6.3 Final remarks

Slang may appear to be a daunting problem domain for natural language processing considering many of its unique characteristics. Despite its ubiquity in daily language use, slang remains under-represented in the literature. By dissecting the problem space and better understanding the underlying semantic principles of slang, I hope my dissertation work makes slang a more accessible research topic for future researchers.

Although this dissertation makes several key contributions towards the automatic processing for slang, it remains an open challenge to design practical NLP systems that are capable of processing slang. Even though my proposed approaches can efficiently learn from dictionary-based data, the performance gap between the processing of slang and conventional language has yet to be closed. State-of-the-art large language models, such as GPT-4 (OpenAI, 2023), have began to reach human performance on many complex tasks (Bubeck et al., 2023) without task-specific fine-tuning. The unprecedented amount of training data from diversified sources also makes slang occurrences much more commonplace, turning slang into a high-resource task. As a result, systems such as ChatGPT become much more capable in processing slang.

While acknowledging these advancements, it is important to note the extent to which high-resource methods such as GPT-4 can be applied. In this dissertation, I considered the usage of a term to be slang if a corresponding entry exists in an

English slang dictionary. In other words, defining what constitutes slang based on lexicographers' professional knowledge. This set of slang, although authoritative, only represents the set of most well-known and well-spread slang. Many slang used in smaller communities may not be represented. For example, the well-studied slang *mans* is used as a first-person, singular reference replacement in Toronto (Denis, 2016). However, none of the slang dictionary resources considered in this dissertation includes it, possibly due to its limited regional spread. The use of slang is also not restricted to English, many cases of which actually involve a mix of different languages (Denis, 2021). As we begin to consider slang used in more niche communities and languages, slang once again becomes a resource-scarce problem, necessitating the use of more data-efficient methods than language modeling. Therefore, once we change perspectives on the kinds of slang we want to handle, many of its core challenges would remain relevant. In the most extreme scenario, I envision personalized AI agents that would attune to slang usages specific to individual households.

The advancement in slang NLP also paves the road for many exciting downstream research opportunities other than artificial intelligence. For example, previous work has shown numerous opportunities in applying dictionary-based NLP techniques for slang to problems such as social media sentimental analysis and find good success (Wu et al., 2018; Aly and van der Haar, 2020; Wilson et al., 2020). I hope that the advancement in automated processing of slang would enable creative solutions to better address difficult problems in related areas such as linguistics and social science.

Appendix A

Resources

A list of resources used in this dissertation to conduct experiments on natural language processing for slang:

Data sources:

- The Online Slang Dictionary (OSD):[1]

  Tasks: Slang generation (Sun et al., 2019, 2021), slang detection (Pei et al., 2019), and slang interpretation Sun et al. (2021).

  Data source: http://onlineslangdictionary.com/

- Green's Dictionary of Slang (GDoS):

  Tasks: Slang generation (Sun et al., 2021) and slang variation (Sun et al., 2022).

  Data source: https://greensdictofslang.com/

- Urban Dictionary (UD):

  Tasks: Slang generation (Sun et al., 2021) and slang interpretation (Ni and Wang, 2017; Sun et al., 2022).

  Data source: https://www.urbandictionary.com/

  Dataset - Kaggle: https://www.kaggle.com/datasets/therohk/

[1] The datasets for both OSD and GDoS cannot be publically distributed due to copyright restrictions. Please contact the corresponding authors for permission to access.

urban-dictionary-words-dataset

Dataset - Ni and Wang (2017): http://www.cs.ucsb.edu/~william/data/slang_ijcnlp.zip

Dataset - Sun et al. (2021): https://github.com/zhewei-sun/slanggen

- Reddit:

  Tasks: Slang variation (Del Tredici and Fernández, 2018; Lucy and Bamman, 2021).

  Data source: https://www.reddit.com/

  Dataset - Lucy and Bamman (2021): https://github.com/lucy3/ingroup_lang

- English Wiktionary:

  Tasks: Slang variation (Sun and Xu, 2022).

  Data source: https://en.wiktionary.org/wiki/English

  Dataset: https://kaikki.org/dictionary/rawdata.html

Code repositories:

- CatGO - Python Library for Categorization (Sun et al., 2019):

  https://github.com/zhewei-sun/catgo

- Slang Generation (Sun et al., 2021):

  https://github.com/zhewei-sun/slanggen

- Slang Interpretation and Translation (Sun et al., 2022):

  https://github.com/zhewei-sun/slanginterp

- Slang Semantic Variation (Sun and Xu, 2022):

  https://github.com/zhewei-sun/slangsemvar

Bibliography

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Eduardo G. Altmann, Janet B. Pierrehumbert, and Adilson E. Motter. 2011. Niche as a determinant of word fate in online groups. *PLOS ONE*, 6(5):1–12.

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.

Asaf Amrami and Yoav Goldberg. 2019. Towards better substitution-based word sense induction. *arXiv*.

Lars-Gunnar Andersson and Peter Trudgill. 1992. *Bad Language*. Penguin Books.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

John Ayto and John Simpson. 2010. *Oxford Dictionary of Modern Slang*. Oxford Paperback Reference. OUP Oxford.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Pierre Baldi and Yves Chauvin. 1993. Neural networks for fingerprint recognition. *Neural Computation*, 5(3):402–418.

David Bamman and Gregory Crane. 2011. Measuring historical word sense variation. New York, NY, USA. Association for Computing Machinery.

David Bamman, Chris Dyer, and Noah A. Smith. 2014a. Distributed representations of geographically situated language. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 828–834, Baltimore, Maryland. Association for Computational Linguistics.

David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014b. Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.

Parminder Bhatia, Robert Guthrie, and Jacob Eisenstein. 2016. Morphological priors for probabilistic neural word embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 490–500, Austin, Texas. Association for Computational Linguistics.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

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.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial disparity in natural language processing: A case study of social media African-American English. *arXiv*.

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.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren E. Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas F. Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, O. Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir P. Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Benjamin Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, J. F. Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Robert Reich, Hongyu Ren, Frieda Rong, Yusuf H. Roohani, Camilo Ruiz, Jack Ryan, Christopher R'e, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishna Parasuram Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei A. Zaharia, Michael

Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the opportunities and risks of foundation models. *ArXiv*.

Jan Botha and Phil Blunsom. 2014. Compositional morphology for word representations and language modelling. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1899–1907, Bejing, China. PMLR.

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.

Virginia Braun and Celia Kitzinger. 2001. "Snatch," "Hole," or "Honey-pot"? Semantic categories and the problem of nonspecificity in female genital slang. *The Journal of Sex Research*, 38(2):146–158.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "Siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, volume 6, pages 737–744. Morgan-Kaufmann.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv*.

Joan L. Bybee, Revere Dale Perkins, and William Pagliuca. 1994. *The evolution of grammar: Tense, aspect, and modality in the languages of the world*, volume 196. University of Chicago Press, Chicago, Illinois.

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.

Frederic G. Cassidy, editor. 1985. *Dictionary of American regional English*. Belknap Press of Harvard University Press, Cambridge, Mass.

Boxing Chen and Colin Cherry. 2014. A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA. Association for Computational Linguistics.

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.

Herbert H. Clark. 1998. Communal lexicons. In *Context in Language Learning and Language Understanding*, pages 63–87. Cambridge University Press.

Paul Cook and Graeme Hirst. 2011. Automatic identification of words with novel but infrequent senses. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 265–274, Singapore. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Paul Cook, Jey Han Lau, Diana McCarthy, and Timothy Baldwin. 2014. Novel word-sense identification. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1624–1635, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Paul Cook, Jey Han Lau, Michael Rundell, Diana McCarthy, and Timothy Baldwin. 2013. A lexicographic appraisal of an automatic approach for detecting new word-senses. In *Kosem et al. (eds.), Proceedings of eLex 2013*, pages 49–65, Tallinn, Estonia.

Paul Cook and Suzanne Stevenson. 2010a. Automatically identifying changes in the semantic orientation of words. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Paul Cook and Suzanne Stevenson. 2010b. Automatically identifying the source words of lexical blends in English. *Computational Linguistics*, 36(1):129–149.

Ryan Cotterell and Hinrich Schütze. 2018. Joint Semantic Synthesis and Morphological Analysis of the Derived Word. *Transactions of the Association for Computational Linguistics*, 6:33–48.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Trans. Speech Lang. Process.*, 4(1).

Tom Dalzell and Eric Partridge. 2009. *The Routledge Dictionary of Modern American Slang and Unconventional English*. Routledge.

Scott Deerwester, Susan T. Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

Marco Del Tredici and Raquel Fernández. 2017. Semantic variation in online communities of practice. In *IWCS 2017 - 12th International Conference on Computational Semantics - Long papers*.

Marco Del Tredici and Raquel Fernández. 2018. The road to success: Assessing the fate of linguistic innovations in online communities. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1591–1603, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Lingjia Deng and Janyce Wiebe. 2015. MPQA 3.0: An entity/event-level sentiment corpus. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328, Denver, Colorado. Association for Computational Linguistics.

Derek Denis. 2016. A note on mans in Toronto. *Toronto Working Papers in Linguistics*, 37.

Derek Denis. 2021. Raptors vs. bucktees: the somali influence on toronto slang*. *Journal of Multilingual and Multicultural Development*, 42(6):565–578.

Aliya Deri and Kevin Knight. 2015. How to make a frenemy: Multitape FSTs for portmanteau generation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–210, Denver, Colorado. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shehzaad Dhuliawala, Diptesh Kanojia, and Pushpak Bhattacharyya. 2016. SlangNet: A WordNet like resource for English slang. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4329–4332, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Bethany K. Dumas and Jonathan Lighter. 1978. Is slang a word for linguists? *American Speech*, 53(1):5–17.

Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407.

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.

Penelope Eckert. 2012. Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual review of Anthropology*, 41:87–100.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2014. Diffusion of lexical change in social media. *PLOS ONE*, 9(11):1–13.

Adam Ek. 2018. Identifying source words of lexical blends in Swedish. In *Proceedings of the The Seventh Swedish Language Technology Conference*, pages 82–85, Stockholm, Sweden".

Jeffrey L. Elman. 1990. Finding structure in time. *Cognitive Science*, 14(2):179–211.

Katrin Erk. 2006. Unknown word sense detection as outlier detection. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pages 128–135, New York City, USA. Association for Computational Linguistics.

Katrin Erk. 2016. What do you know about an alligator when you know the company it keeps? *Semantics and Pragmatics*, 9(17):1–63.

Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. 2009. Unsupervised Type and Token Identification of Idiomatic Expressions. *Computational Linguistics*, 35(1):61–103.

Renato Ferreira Pinto Jr. and Yang Xu. 2021. A computational theory of child overextension. *Cognition*, 206:104472.

John R. Firth. 1957. *Papers in Linguistics, 1934-1951*. Oxford University Press.

Stuart Berg Flexner. 1960. *Dictionary of American slang*. Thomas Y. Crowell Company.

Lea Frermann and Mirella Lapata. 2016. A Bayesian model of diachronic meaning change. *Transactions of the Association for Computational Linguistics*, 4:31–45.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Varun Gangal, Harsh Jhamtani, Graham Neubig, Eduard Hovy, and Eric Nyberg. 2017. Charmanteau: Character embedding models for portmanteau creation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2917–2922, Copenhagen, Denmark. Association for Computational Linguistics.

Eric Gilbert and Karrie Karahalios. 2009. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, page 211–220, New York, NY, USA. Association for Computing Machinery.

David Goldberg, David Nichols, Brian M. Oki, and Douglas Terry. 1992. Using collaborative filtering to weave an information tapestry. *Commun. ACM*, 35:61–70.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. In *International Conference of Machine Learning (ICML) 2012 Workshop on Representation Learning*.

Jonathan Green. 2010. *Green's Dictionary of Slang*. Chambers, London.

Karan Grewal and Yang Xu. 2021. Chaining algorithms and historical adjective extension. *Computational approaches to semantic change (Language Variation)*, 6:189.

Kristina Gulordava and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books ngram corpus. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh, UK. Association for Computational Linguistics.

Anshita Gupta, Sanya Bathla Taneja, Garima Malik, Sonakshi Vij, Devendra K. Tayal, and Amita Jain. 2019. Slangzy: a fuzzy logic-based algorithm for english slang meaning selection. *Progress in Artificial Intelligence*, 8(1):111–121.

Amir Ahmad Habibi, Charles Kemp, and Yang Xu. 2020. Chaining and the growth of linguistic categories. *Cognition*, 202:104323.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Eduard H. Hovy. 1990. Pragmatics and natural language generation. *Artificial Intelligence*, 43(2):153–197.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. Association for Computational Linguistics.

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.

Andres Karjus, Richard A. Blythe, Simon Kirby, Tianyu Wang, and Kenny Smith. 2021. Conceptual similarity and communicative need shape colexification: An experimental study. *Cognitive Science*, 45(9):e13035.

Daphna Keidar, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. Slangvolution: A causal analysis of semantic change and frequency dynamics in slang. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1422–1442, Dublin, Ireland. Association for Computational Linguistics.

Charles Kemp and Terry Regier. 2012. Kinship categories across languages reflect general communicative principles. *Science*, 336(6084):1049–1054.

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*.

Devorah E. Klein and Gregory L. Murphy. 2001. The representation of polysemous words. *Journal of Memory and Language*, 45(2):259–282.

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.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Vivek Kulkarni, Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2015. Statistically significant detection of linguistic change. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 625–635, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Vivek Kulkarni and William Yang Wang. 2017. Tfw, damngina, juvie, and hotsietotsie: On the linguistic and social aspects of internet slang. *arXiv*.

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.

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.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

George Lakoff. 1987. *Women, fire, and dangerous things: what categories reveal about the mind*. University of Chicago Press.

Sidney Landau. 1984. *Dictionaries: The art and craft of lexicography*. Charles Scribner's Sons, New York, NY.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 591–601, Avignon, France. Association for Computational Linguistics.

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.

Adrienne Lehrer. 1985. The influence of semantic fields on semantic change. *Historical Semantics: Historical Word Formation*, 29:283–296.

Esther Lewin and Albert E. Lewin. 1988. *The Random House Thesaurus of Slang: 150,000 Uncensored Contemporary Slang Terms, Common Idioms, and Colloquialisms Arranged for Quick and Easy Reference*. Random House.

Wang Ling, Chris Dyer, Alan W Black, Isabel Trancoso, Ramón Fermandez, Silvio Amir, Luís Marujo, and Tiago Luís. 2015. Finding function in form: Compositional character models for open vocabulary word representation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1520–1530, Lisbon, Portugal. Association for Computational Linguistics.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv*.

Li Lucy and David Bamman. 2021. Characterizing English Variation across Social Media Communities with BERT. *Transactions of the Association for Computational Linguistics*, 9:538–556.

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.

Barbara C. Malt, Steven A. Sloman, Silvia Gennari, Meiyi Shi, and Yuan Wang. 1999. Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, 40:230–262.

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

Elisa Mattiello. 2009. Difficulty of slang translation. In *Translation Practices*, pages 65–83. Brill Rodopi.

Merriam-Webster, editor. 2004. *Merriam-Webster's collegiate dictionary*. Merriam-Webster.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Sunny Mitra, Ritwik Mitra, Suman Kalyan Maity, Martin Riedl, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2015. An automatic approach to identify word sense changes in text media across timescales. *Natural Language Engineering*, 21(5):773–798.

Sunny Mitra, Ritwik Mitra, Martin Riedl, Chris Biemann, Animesh Mukherjee, and Pawan Goyal. 2014. That's sick dude!: Automatic identification of word sense change across different timescales. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1020–1029, Baltimore, Maryland. Association for Computational Linguistics.

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.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational Sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Dong Nguyen, Barbara McGillivray, and Taha Yasseri. 2018. Emo, love and god: making sense of urban dictionary, a crowd-sourced online dictionary. *Royal Society open science*, 5(5):172320.

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.

Robert M. Nosofsky. 1986. Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115:39–57.

OpenAI. 2023. GPT-4 technical report. *arXiv*.

Alok Ranjan Pal and Diganta Saha. 2013. Detection of slang words in e-data using semi-supervised learning. *International Journal of Artificial Intelligence and Applications*, 4(5):49–61.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. ACL '02, page 311–318, USA. Association for Computational Linguistics.

Hermann Paul. 1880. *Prinzipien der Sprachgeschichte*. Niemeyer (Halle a. S.).

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.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71.

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.

Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. 2017. Mimicking word embeddings using subword RNNs. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 102–112, Copenhagen, Denmark. Association for Computational Linguistics.

Yuval Pinter, Cassandra L. Jacobs, and Jacob Eisenstein. 2020. Will it unblend? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1525–1535, Online. Association for Computational Linguistics.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

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.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

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.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104:192–233.

Maria Ryskina, Ella Rabinovich, Taylor Berg-Kirkpatrick, David Mortensen, and Yulia Tsvetkov. 2020. Where new words are born: Distributional semantic analysis of neologisms and their semantic neighborhoods. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 367–376, New York, New York. Association for Computational Linguistics.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 104–111, Athens, Greece. Association for Computational Linguistics.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5149–5152.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Haoyue Shi, Luke Zettlemoyer, and Sida I. Wang. 2021. Bilingual lexicon induction via unsupervised bitext construction and word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 813–826, Online. Association for Computational Linguistics.

Steven A. Sloman, Barbara C. Malt, and Arthur Fridman. 2001. Categorization versus similarity: The case of container names. *Similarity and categorization*, page 73–86.

James Slotta. 2016. Slang and the semantic sense of identity. *UT Faculty/Researcher Works*.

Jake Snell, Kevin Swersky, and Richard 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.

Karl Sornig. 1981. *Lexical Innovation: A Study of Slang, Colloquialisms, and Casual Speech*. John Benjamins B.V., Amsterdam.

Richard A Spears. 1981. *Slang and euphemism*. Signet Book.

Gustaf Stern. 1931. *Meaning and change of meaning; with special reference to the English language*. Wettergren & Kerbers.

Angus Stevenson, editor. 2010. *Oxford dictionary of English*. Oxford University Press, USA.

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.

Zhewei Sun and Yang Xu. 2022. Tracing semantic variation in slang. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1299–1313, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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.

Zhewei Sun, Richard Zemel, and Yang Xu. 2021. A Computational Framework for Slang Generation. *Transactions of the Association for Computational Linguistics*, 9:462–478.

Zhewei Sun, Richard Zemel, and Yang Xu. 2022. Semantically informed slang interpretation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5213–5231, Seattle, United States. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27, pages 3104–3112. Curran Associates, Inc.

Bradley A. Swerdfeger. 2012. Assessing the viability of the urban dictionary as a resource for slang.

Sali A. Tagliamonte and Derek Denis. 2008. Linguistic ruin? LOL! instant messaging and teen language. *American speech*, 83(1):3–34.

Neil C. Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. 2020. The computational limits of deep learning. *arXiv*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Stephen Ullmann. 1942. The range and mechanism of changes of meaning. *The Journal of English and Germanic Philology*, 41(1):46–52.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, page 6000–6010. Curran Associates, Inc.

Tony Veale, Ekaterina Shutova, and Beata Beigman Klebanov. 2016. *Metaphor: A computational perspective*. Morgan & Claypool Publishers.

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.

Isaac Waller and Ashton Anderson. 2020. Community embeddings reveal large-scale cultural organization of online platforms. *arXiv*.

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.

Beatrice. Warren. 1992. *Sense Developments: A Contrastive Study of the Development of Slang Senses and Novel Standard Senses in English*. Acta Universitatis Stockholmiensis: Stockholm studies in English. Almqvist & Wiksell International.

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.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Charagram: Embedding words and sentences via character n-grams. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1515, Austin, Texas. Association for Computational Linguistics.

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.

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.

Liang Wu, Fred Morstatter, and Huan Liu. 2018. SlangSD: Building, expanding and using a sentiment dictionary of slang words for short-text sentiment classification. *Lang. Resour. Eval.*, 52(3):839–852.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yang Xu and Charles Kemp. 2015. A computational evaluation of two laws of semantic change. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, pages 2703–2708. Cognitive Science Society.

Yang Xu, Terry Regier, and Barbara C Malt. 2016. Historical semantic chaining and efficient communication: The case of container names. *Cognitive science*, 40(8):2081–2094.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R. Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Chuanrun Yi, Dong Wang, Chunyu He, and Ying Sha. 2019. Learning to explain chinese slang words. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series*, pages 22–33, Cham. Springer International Publishing.

Tatu Ylonen. 2022. Wiktextract: Wiktionary as machine-readable structured data. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 1317–1325, Marseille, France.

Lei Yu and Yang Xu. 2021. Predicting emergent linguistic compositions through time: Syntactic frame extension via multimodal chaining. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 920–931, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lotfi A. Zadeh. 1965. Fuzzy sets. *Information and Control*, 8(3):338–353.

Ziheng Zeng and Suma Bhat. 2021. Idiomatic expression identification using semantic compatibility. *Transactions of the Association for Computational Linguistics*, 9:1546–1562.

Ziheng Zeng and Suma Bhat. 2022. Getting BART to Ride the Idiomatic Train: Learning to Represent Idiomatic Expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137.

Wanzheng Zhu, Hongyu Gong, Rohan Bansal, Zachary Weinberg, Nicolas Christin, Giulia Fanti, and Suma Bhat. 2021. Self-supervised euphemism detection and identification for content moderation. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 229–246.