Semantically Informed Slang Interpretation

Zhewei Sun[1], **Richard Zemel**[1,2,4], **Yang Xu**[1,3,4]

[1]Department of Computer Science, University of Toronto, Toronto, Canada
[2]Department of Computer Science, Columbia University, New York, USA
[3]Cognitive Science Program, University of Toronto, Toronto, Canada
[4]Vector Institute for Artificial Intelligence, Toronto, Canada
{zheweisun, zemel, yangxu}@cs.toronto.edu

Abstract

Slang is a predominant form of informal language making flexible and extended use of words that is notoriously hard for natural language processing systems to interpret. Existing approaches to slang interpretation tend to rely on context but ignore semantic extensions common in slang word usage. We propose a semantically informed slang interpretation (SSI) framework that considers jointly the contextual and semantic appropriateness of a candidate interpretation for a query slang. We perform rigorous evaluation on two large-scale online slang dictionaries and show that our approach not only achieves state-of-the-art accuracy for slang interpretation in English, but also does so in zero-shot and few-shot scenarios where training data is sparse. Furthermore, we show how the same framework can be applied to enhancing machine translation of slang from English to other languages. Our work creates opportunities for the automated interpretation and translation of informal language.

1 Introduction

Slang is one of the most common forms of informal language, but interpreting 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, slang interpretation is also notoriously difficult for state-of-the-art natural language processing (NLP) systems, which presents a critical challenge to downstream applications such as natural language understanding and machine translation.

Consider the sentence "I got really *steamed* when my car broke down". As illustrated in Figure 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

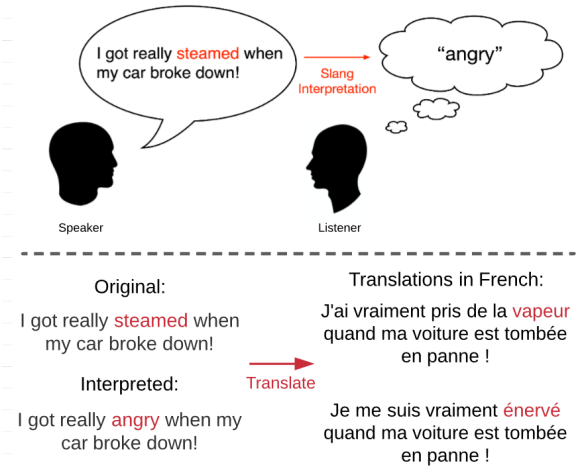


Figure 1: Illustrations of slang interpretation in English (top panel) and slang translation (bottom panel) from English to French on the original sentence (nonsensical), or on the interpreted version of the sentence (sensical).

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 1 shows the system would be quite effective in translating the intended meaning.

Here we consider the problem of slang interpretation illustrated in the top panel of Figure 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 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

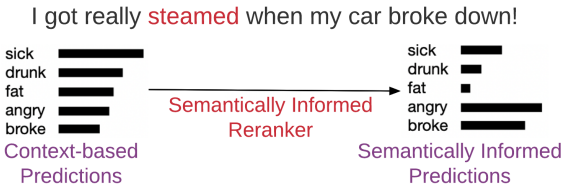


Figure 2: Workflow of the proposed framework.

be considered a candidate paraphrase for the target slang *steamed* conditioned on its surrounding words. Although the groundtruth 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-contained 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 recent work on generative models of slang (Sun et al., 2019, 2021), 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 (Sun et al., 2021). 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.

Figure 2 illustrates the workflow of our approach. 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.

For the scope of this work, we focus on interpreting slang expressions with existing word forms because 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 state-of-the-art language models in slang interpretation in English and slang translation from English to other languages.[1]

2 Related Work

2.1 Natural Language Processing for Slang

Existing approaches in the natural language processing for slang focus on efficient construction, extension, and retrieval from dictionary-based resources for detection (Pal and Saha, 2013; Dhuliawala et al., 2016), interpretation (Gupta et al., 2019), and sentiment analysis of slang (Dhuliawala et al., 2016; Wu et al., 2018). These studies often rely on heuristic measures to determine or retrieve the meaning of slang and cannot generalize beyond what was available in the training data. Recent work such as Kulkarni and Wang (2018) and Pei et al. (2019) proposed deep learning based approaches to generalize toward unseen slang.

Closely related to our study is Ni and Wang (2017) that 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 groundtruth slang definition (also in English), and such pairs are fed into a translation model. In addition, the spellings of slang word forms are also considered as input. In their model, both the context and the slang form are encoded using separate LSTM encoders. 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

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

time, beam search (Graves, 2012) is applied to decode a set of candidate definition sentences.

One key problem with this approach is that the Dual Encoder 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.

2.2 Generative Semantic Models of Slang

Recent work by Sun et al. (2019, 2021) proposed a neural-probabilistic generative framework for modeling slang word choice in novel context. Given a query sentence with the target slang blanked out and the intended meaning of that slang, their framework predicts which word(s) would be appropriate slang choices that fill in the blank. Relevant to their framework is a semantic model of slang that uses contrastive learning from Siamese networks (Baldi and Chauvin, 1993; Bromley et al., 1994) to relate conventional and slang meanings of words. This model yields a semantic embedding space that is sensitive to flexible slang meaning extensions. For example, it may learn that meanings associated with "vapor" can extend to meanings about "angry" (as in the *steamed* example in Figure 1).

Differing from slang generation, our work concerns the inverse problem of slang interpretation that has more direct applications in natural language processing particularly machine translation (e.g., of informal language). Building on work of slang generation, we incorporate the generative semantic model of slang in a semantically informed interpretation framework that integrates context to infer the intended meaning of a target slang.

3 Computational Framework

Our computational framework is comprised of three key components following the workflow illustrated in Figure 2: 1) A context-based baseline interpreter that generates an n-best list of candidate interpretations for a target slang in a query sentence; 2) A semantic model of slang that checks the appropriateness of a candidate interpretation to the slang context; 3) A reranker informed by the semantic model in 2) that re-prioritizes the candidate interpretations from the context-based interpreter in 1). We use this framework for both interpreting slang within English and translating slang from English to other languages.

3.1 Context-based Interpretation

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. 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:

$$I(S, C_S) = \arg\max_{M} P(M|S, C_S) \tag{1}$$

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 alternative 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).

LM-based interpreter. 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:

$$\begin{aligned} I(S, C_S) = & D[\arg\max_{w} LM(w|C_S \backslash S) \\ & + \mathbb{1}_{T(w)}[T(C_S \backslash S)]] \end{aligned} \tag{2}$$

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. 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. 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.

Dual encoder. 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. 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.

While this approach is trained directly on slang data and considers the slang word forms, it requires a large on-task dataset to be trained effectively. This model also does not take into account the appropriateness of meaning extension in slang usage. We next describe how a semantic model of slang can be incorporated to enhance the context-based interpreters.

3.2 Semantic Model of Slang

Given an n-best list of candidate interpretations $\mathcal{K}$ for the target slang S in context C_S, 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. Sun et al. (2019, 2021) modeled 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:

$$\begin{aligned} f(k, S) &= sim(E_k, E_S) \\ &= \exp(-\frac{d(E_k, E_S)}{h_m}) \end{aligned} \tag{3}$$

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 ($\mathcal{E}_S$):

$$E_S = \frac{1}{|\mathcal{E}_S|} \sum_{E_{S_i} \in \mathcal{E}_S} E_{S_i} \tag{4}$$

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

Following Sun et al. (2021), 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:

$$Loss = \Big[d(E_{SL}, E_P) - d(E_{SL}, E_N) + m\Big]_+ \tag{5}$$

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.

3.3 Semantically 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):

$$SSI(\mathcal{K}) = \arg\max g(k, S) \tag{6}$$

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 3:

$$g(k, S) = P(k|S) \propto f(k, S) \quad (7)$$

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:

$$g^*(k, S) \propto \sum_{S' \in L(S)} sim(S, S') g(k, S') \quad (8)$$

$$sim(S, S') = \exp(-\frac{d(S, S')}{h_{cf}}) \quad (9)$$

Here, $d(S, S')$ is the cosine distance between the two slang's word vectors 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 (Sun et al., 2019, 2021).

4 Datasets

We use two online English slang dictionary resources to train and evaluate our proposed slang interpretation framework: 1) the Online Slang Dictionary (OSD)[2] dataset from Sun et al. (2021) and 2) a collection of Urban Dictionary (UD)[3] 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 example 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[4], 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 exact references 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 1 shows the size of the datasets after pre-processing.

[2]OSD: http://onlineslangdictionary.com
[3]UD: https://www.urbandictionary.com
[4]OD: https://en.oxforddictionaries.com

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.

5 Evaluation and Results

5.1 Evaluation on Slang Interpretation

We first evaluate the semantically informed and baseline interpretation models in a multiple choice task. In this task, each query is paired with a set of definitions that construe the meaning of the target slang in the query. One of these definitions is the groundtruth meaning of the target slang, while the other definitions are incorrect or negative entries sampled from the training set (i.e., all taken from the slang dictionary resources described). 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 groundtruth 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 groundtruth 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 groundtruth 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 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

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 1: Summary of basic statistics for the two online slang dictionaries used in the study.

Model	Distinctively sampled candidates	Randomly sampled candidates
Dataset 1: Online Slang Dictionary (OSD) (Sun et al., 2021)		
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 2: Evaluation of English slang interpretation measured in mean-reciprocal rank (MRR). Predictions are ranked against 4 negative candidates distinctively or randomly sampled, yielding MRR=0.457 for the random baseline.

entries corresponding to words found in the OSD testset are filtered out in this particular experiment. Detailed training procedures for all models can be found in Appendix A.

Table 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.

Table 3 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 groundtruth 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

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.
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.

Table 3: Example queries from OSD and top predictions made from both the baseline language infill models (LM Infill) and the Dual Encoder models with $n = 50$, along with top predictions from the enhanced semantically informed slang interpretation (SSI) models. Additional examples can be found in Appendix B.1.

the two datasets might have caused the difference in performance gain.

5.2 Zero-shot and Few-shot 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 finetune the language infill model (LM Infill) on the first example usage sentence that correspond 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 finetune 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 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 is only trained on entries from the training set.

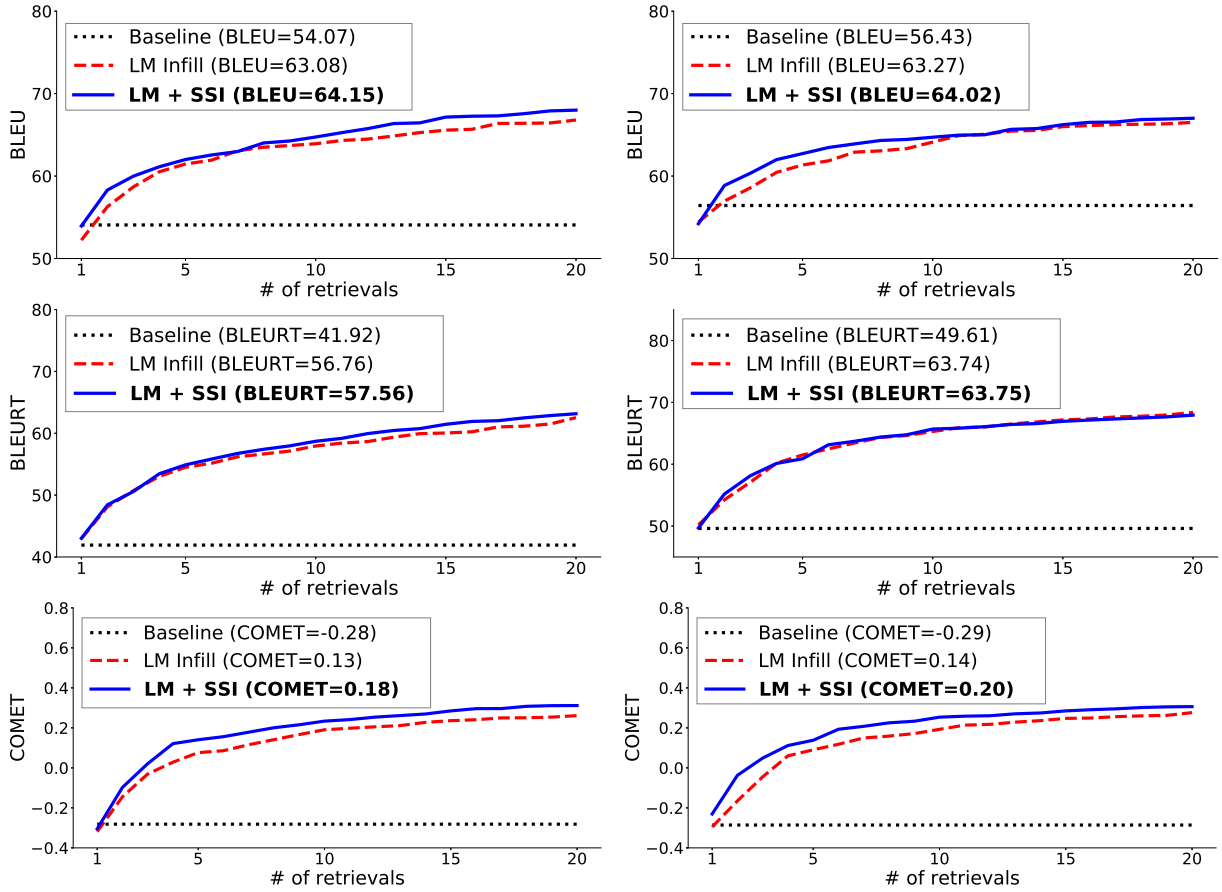
Model	Distinct negatives	Random negatives
LM Zero-shot, $n = 50$	0.444	0.443
+ SSI	**0.571**	**0.565**
LM Few-shot, $n = 50$	0.504	0.513
+ SSI	**0.567**	**0.564**

Table 4: Interpretation results on OSD measured in mean-reciprocal rank (MRR) before and after finetuning the language infill model.

5.3 Evaluation on 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 2), which allows direct insertion of the resulting interpretation into the original sentence.

We perform our experiment on the OSD test set because it contains higher quality example sentences than UD. To mitigate potential biases, we



(a) English to French

(b) English to German

Figure 3: 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.

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 gold-standard 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.

Figure 3 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 groundtruth definitions, the set of interpreted sentences often contain

Query (target slang in ***bold italic***):	I want to go get coffee but it's ***bitter*** outside.
Definition of target slang:	Abbreviated form of bitterly cold.
Groundtruth 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)
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)

Table 5: An example of machine translation of slang, without or with the application of the SSI framework. The top 2 interpreted and translated sentences are shown for each model with BLEU, BLEURT, and COMET scores against the gold-standard translation shown in parentheses. More examples can be found in Appendix B.4.

plausible interpretations that result in improved translation of slang. Table 5 provides 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. Our approach may be ultimately integrated with a slang detector (e.g., Pei et al. 2019) to produce fully automated translations in natural context that involves slang.

6 Conclusion

The flexible nature of slang is a hallmark of informal language, and to our knowledge 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 in this area may benefit from principled approaches that model the coinage of slang expressions with novel word forms and multi-word expressions with complex formation strategies, as well as how slang terms emerge in specific individuals and groups. Our current study shows promise for advancing methodologies in informal language processing toward these avenues of future research.

Ethical Considerations

We analyze entries of slang usage in our work and acknowledge that such usages may contain offensive information. We retain such entries in our datasets to preserve the scientific validity of our results, as a significant portion of slang usage aligns to possibly offensive usage context. In the presentation our of results, however, we strive to select examples or illustrations that minimize the extent to which offensive content is represented. We also acknowledge that models trained on datasets such as the Urban Dictionary have a greater tendency to generate offensive language. All model outputs shown are results of model learning and do not reflect opinions of the authors and their affiliated organizations. We hope that our work will contribute to the greater good by enhancing AI system's ability to comprehend such offensive language use, allowing better filtering of online content that may be potentially harmful.

Acknowledgements

We thank the ARR reviewers for their constructive comments and suggestions, and Walter Rader for permission to use The Online Slang Dictionary. This work was supported by a NSERC Discovery Grant RGPIN-2018-05872, a SSHRC Insight Grant #435190272, and an Ontario ERA Award to YX.

References

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.

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

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

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.

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. 1994. Signature verification using a "siamese" time delay neural network. In J. D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, pages 737–744. Morgan-Kaufmann.

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.

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.

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.

Connie C Eble. 2012. *Slang & sociability: In-group language among college students*. University of North Carolina Press, Chapel Hill, NC.

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.

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.

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.

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*.

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.

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

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

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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

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.

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.

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.

A Training Procedures

A.1 Baseline Models

We train two context-based slang interpreters described in Section 3.1 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 2, if a matching dictionary entry can be found in Oxford Dictionary (OD), the top 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 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 took 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.

A.2 Semantic Reranker

We obtain the contrastive sense encodings (CSE) described in Section 3.2 by using 768-dimensional Sentence-BERT (Reimers and Gurevych, 2019) embeddings as our baseline embedding. Following Sun et al. (2021), we train the contrastive network with a 1.0 margin (m in Equation 5) using Adam (Kingma and Ba, 2015) with a learning rate of 2^{-5}, resulting in 768-dimensional definition sense presentations. 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 Sun et al. (2021) to estimate the kernel width parameters (h_m in Equation 3 and h_{cf} in Equation 9) 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 9 by comparing their fastText (Bojanowski et al., 2017) embeddings. For collaborative filtering, the neighborhood of words $L(S)$ in Equation 8 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.

B Additional Results

B.1 Additional Interpretation Examples

Table 7 show additional example interpretations made by the models evaluated in Section 5.1. The first three examples illustrate cases where the semantically informed models were not able to predict the exact definitions, but came up with definitions that are more closely related to the groundtruth compared to the baseline. The latter two examples show cases where the semantically informed models fail to make an improvement.

B.2 Effect of Context Length

In the model evaluation described in Section 5.1, 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 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

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 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.

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.

B.3 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 finetuned results to simulate a stronger baseline model. Table 6 shows the results. When compared to the zero-shot results in Table 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 improved interpretation candidates from the finetuned 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.

B.4 Machine Translation Examples

Table 8 to Table 11 show full example translations (English to French) made for the experiment described in Section 5.3, translating sentences containing slang before and after applying slang interpretation.

C Data Permissions

At the time when the research is performed, Online Slang Dictionary (OSD) explicitly forbids automated downloading of data from its website service. We therefore have obtained written permission from its owner to download and use the dataset for personal research use. We download data from the online version of the Oxford Dictionary (OD) under personal use. We cannot publically share the two datasets used above as a result. Readers interested in obtaining the exact datasets used in this work must first obtain relevant permission from the respective data owner before the authors of this work can share the data. The Urban Dictionary (UD) dataset is obtained from the authors of Ni and Wang (2017) under a research only license. We release entries relevant to our study with the original data license attached.

[Example 1]	
Query (target slang in ***bold italic***):	That girl has a ***donkey***.
Groundtruth 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 2]	
Query:	I am an ***onion***.
Groundtruth 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 3]	
Query:	In Blastem version 4, they really ***nerf*** the EnemyToaster.
Groundtruth 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.
[Example 4]	
Query:	I heard Steve was sent to the ***cooler*** for breaking and entering.
Groundtruth 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 5]	
Query:	Do you have any ***safety***
Groundtruth 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 7: Additional examples: 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.

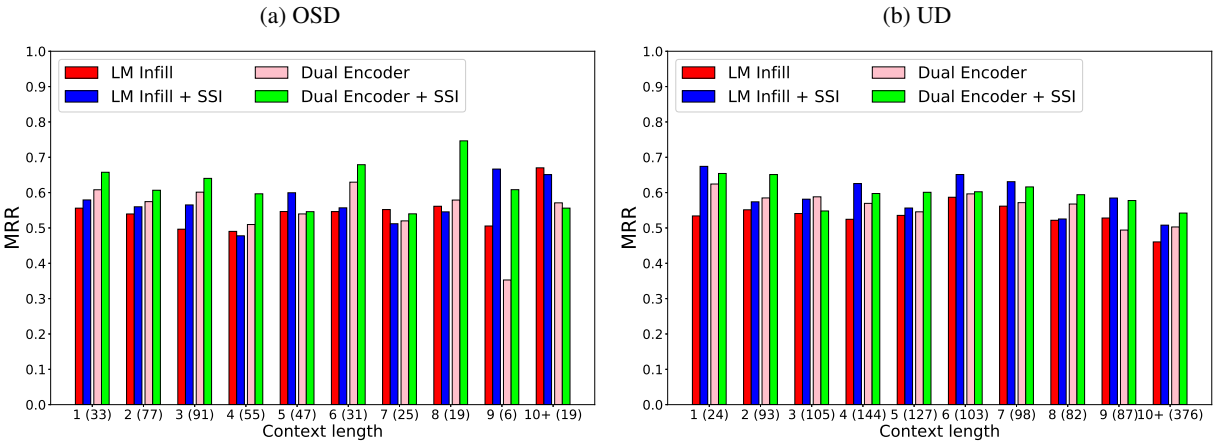


Figure 4: Evaluation of slang interpretation performance measured in mean-reciprocal rank (MRR) for all models with $n = 50$. Test entries are partitioned based on the number of content words (excluding the slang expression itself) found within the corresponding example usage sentence. Number of entries corresponding to each context length is shown in parenthesis on the x-axis legend.

[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.
Groundtruth 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 8: Additional 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.

[Example 2]	
Query:	That band was so totally ***vast***.
Definition of target slang:	Cool or anything good.
Groundtruth 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 9: Continuation of Table 8.

[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.
Groundtruth 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 10: Continuation of Table 9.

[Example 4]	
Query:	I want to go get coffee but it's ***bitter*** outside.
Definition of target slang:	Abbreviated form of bitterly cold.
Groundtruth 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 *coule* 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 11: Continuation of Table 10.