Word Sense Extension

Lei Yu[1], **Yang Xu**[1,2]
[1] Department of Computer Science, University of Toronto
[2] Cognitive Science Program, University of Toronto
{jadeleiyu,yangxu}@cs.toronto.edu

Abstract

Humans often make creative use of words to express novel senses. A long-standing effort in natural language processing has been focusing on word sense disambiguation (WSD), but little has been explored about how the sense inventory of a word may be extended toward novel meanings. We present a paradigm of *word sense extension* (WSE) that enables words to spawn new senses toward novel context. We develop a framework that simulates novel word sense extension by first partitioning a polysemous word type into two pseudo-tokens that mark its different senses, and then inferring whether the meaning of a pseudo-token can be extended to convey the sense denoted by the token partitioned from the same word type. Our framework combines cognitive models of chaining with a learning scheme that transforms a language model embedding space to support various types of word sense extension. We evaluate our framework against several competitive baselines and show that it is superior in predicting plausible novel senses for over 7,500 English words. Furthermore, we show that our WSE framework improves performance over a range of transformer-based WSD models in predicting rare word senses with few or zero mentions in the training data.

1 Introduction

Humans make creative reuse of words to express novel senses. For example, the English verb *arrive* extended from its original sense "to come to locations (e.g., to *arrive* at the gate)" toward new senses such as "to come to an event (e.g., to *arrive* at a concert)" and "to achieve a goal or cognitive state (e.g., to *arrive* at a conclusion)" (see Figure 1). The extension of word meaning toward new context may draw on different cognitive processes such as metonymy and metaphor, and here we develop a general framework that infers how words extend to plausible new senses.

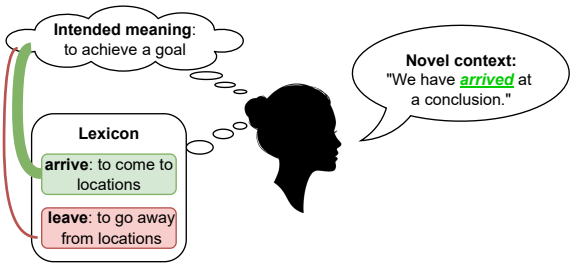

Figure 1: Illustration of the problem of word sense extension. Given a novel context, a speaker chooses an existing word in the lexicon to convey a novel intended meaning that has not appeared in the semantics of that word. The speaker determines the appropriateness of a chosen word (indicated by line width of the colored curves) based on semantic relatedness between the novel intended meaning and existing word meanings.

A long-standing effort in natural language processing (NLP) is to build systems that support automatic word sense disambiguation (WSD) from linguistic context. This line of work typically takes a discriminative approach toward word meaning and has developed models relying on both traditional machine learning (Gale et al., 1992; Kilgarriff and Rosenzweig, 2000; Zhong and Ng, 2010; Iacobacci et al., 2016) and modern neural language models (Huang et al., 2019; Wiedemann et al., 2019; Loureiro and Jorge, 2019; Bevilacqua and Navigli, 2020). However, existing WSD models often struggle with recognizing rare word senses with few or no mentions in training (Blevins et al., 2021). Here we show that by modelling the generative extensional processes of word meaning, WSD models can become better at recognizing infrequent word senses in natural context and without relying on external lexical resources.

Work in computational and cognitive linguistics shows that word senses do not extend arbitrarily (Nunberg, 1979; Lehrer, 1990; Rumshisky and Batiukova, 2008). Lexical semanticists have suggested that a number of cognitive devices may be

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3281–3294

applied to generate creative word usages, such as logical metonymy (Copestake and Briscoe, 1995; Pustejovsky, 1998) and metaphor (Lakoff and Johnson, 2008; Pustejovsky and Rumshisky, 2010). Cognitive linguists have also suggested that systematic mappings between conceptual domains underlie the metaphorization of word meaning (Brugman and Lakoff, 1988; Lakoff and Johnson, 2008; Gentner, 1983). However, the reliance on hand-crafted rules of semantic productivity makes it difficult to implement systems that support flexible and scalable extension to new word senses.

We present a paradigm that considers the problem of *word sense extension* (WSE) illustrated in Figure 1. Given a novel context and an intended meaning, a speaker wishes to choose an existing word in the lexicon to express that meaning which the word has never been used to convey. To operationalize a speaker model without prior knowledge about pairings between the novel meaning and existing word forms, we replace each candidate word type with a pair of "pseudo-tokens" that signify one of its existing senses (called the target sense) the other senses (called the source senses) respectively, a method related to previous work in polysemy induction (Pilehvar and Navigli, 2014; Dubossarsky et al., 2018). We then infer whether a partitioned pseudo-token denoting the source sense may be extended to express the target sense denoted by its sibling token partitioned from the same word type. We propose a family of cognitively-inspired probabilistic models for this inference problem. We show that our WSE models can reliably predict plausible novel senses on a large usage-based dataset with approximately 34,000 senses for over 7,500 English word types.[1]

2 Related work

2.1 Models of word meaning extension

Researchers in lexical semantics and cognitive linguistics have both proposed theories to account for the malleable nature of lexical meaning. The Generative Lexicon theory by Pustejovsky (1998) argues that a fixed set of generative devices, such as type-coercion and co-composition, can operate on the lexical structure a word to produce various related meaning interpretations. Copestake and Briscoe (1995) also illustrates how formal lexical rules such as grinding and portioning can be applied to produce novel word usages such as logical metonymy. In cognitive linguistics, Lakoff (1987) argues that word meanings grow relying on processes of chaining, whereby novel meanings link to existing ones that are close in semantic space. Similar processes are also relevant to the construction of metaphorical usages in natural language drawing on image schemas (Brugman and Lakoff, 1988; Dewell, 1994; Gibbs Jr and Colston, 2008) and analogy or structural alignment between domains (Gentner, 1983; Falkenhainer et al., 1989).

Our work builds on the cognitive theory and recent computational work on chaining (Lakoff, 1987; Malt et al., 1999; Ramiro et al., 2018; Habibi et al., 2020; Grewal and Xu, 2020; Yu and Xu, 2021), and we show that a chaining-based framework learns systematic patterns of word sense extension discussed in the tradition of generative lexical semantics. Related work has taken a similar approach for modelling sense extension in slang usages (Sun et al., 2021), but here we consider the more general problem of word sense extension.

2.2 Models of word sense disambiguation

A large community in NLP has been working on the problem of word sense disambiguation (WSD). Early WSD systems adopt a knowledge-based approach by comparing the neighborhood context of a target word with its gloss or definition in lexicographic databases such as WordNet (Miller, 1995; Gale et al., 1992; Kilgarriff and Rosenzweig, 2000). Later work develops feature-based classification models to predict sense labels for a word based on its linguistic features (Zhong and Ng, 2010; Iacobacci et al., 2016; Raganato et al., 2017). Recent progress in deep learning also motivates the development of WSD systems based on deep contextualized language models (CLM) or its combination with external lexical knowledge base (Huang et al., 2019; Hadiwinoto et al., 2019; Bevilacqua and Navigli, 2020). Despite these impressive advances, many CLM-based WSD systems still suffer from the data sparsity that stems from the Zipfian distribution of word senses (Kilgarriff, 2004) – i.e. the most frequent sense of a polysemous word often accounts for a dominant portion of its mentions, while other senses have much less or even zero frequency in training data. Recent work has proposed to mitigate this sense sparsity problem by resorting to gloss information (Luo et al., 2018; Kumar et al.,

[1]We release the code and data for our work here: https://github.com/jadeleiyu/word_sense_extension.

2019; Huang et al., 2019; Blevins and Zettlemoyer, 2020) or non-parametric few-shot learning (Holla et al., 2020; Chen et al., 2021). We shall demonstrate that learning word sense extensions offers an alternative approach to improve WSD system performance on infrequent word senses by leveraging the systematic semantic relational patterns between conventional and novel word senses.

2.3 Contextualized semantic representations

Existing work has proposed to apply contextualized language models to lexical semantic tasks that involve polysemy. Diachronic studies show that contextualized representations of word usage and sense definitions can be used to detect lexical semantic shifts (Giulianelli et al., 2020; Hu et al., 2019). Probing studies also suggest that pretrained contextualized language models encode rich lexical semantic information that may help decide the levels of word polysemy (Garí Soler and Apidianaki, 2021) and infer semantic relations between word senses (Vulić et al., 2020). The WSE paradigm we propose is related to lexical substitution, where a model is used to replace a target word in a sentence with a substitute word without changing the sentence meaning (McCarthy and Navigli, 2007; Melamud et al., 2016; Zhou et al., 2019). However, our framework goes beyond this research by asking whether a word can extend its sense inventory to express novel intended meanings in natural context.

3 Computational framework

Our framework of word sense extension involves three interrelated components: 1) A procedure for partitioning polysemous words in the lexicon into new pseudo-tokens that signify their different senses; 2) a probabilistic, chaining-based formulation of word sense extension for lexical choice making under novel linguistic context; and 3) a learning algorithm for a transformed semantic space to learn flexible extensions of word senses.

3.1 Sense-based word type partitioning

Let $\mathcal{W} = \{w_1, ..., w_{|V|}\}$ be our vocabulary of polysemous (English) word types, where each w has a set of n senses $\mathcal{S}_w = \{s_1, ..., s_n\}$. Assume that for each w there is also a collection of its sense-annotated sample usage contexts $\mathcal{C}_w = \{(c_1, y_1), ..., (c_m, y_m)\}$, where each contextual sequence $c \in \mathcal{C}_w$ is labeled with a sense $y \in \mathcal{S}_w$ instantiating the meaning of w in that usage context. We want to simulate the scenario where a speaker, without knowing a priori that a word w has a sense $s^* \in \mathcal{S}_w$, is able to extend the meaning of w to expressing s under novel context.

To operationalize this idea of word sense extension, we first partition each w into two hypothetical tokens: a source token t^0 that denotes the set of existing source senses $\mathcal{S}_0 = \mathcal{S} \setminus \{s\}$ of w, and a target token t^* that denotes the novel target sense s^* to which w extends beyond its existing senses. We then replace w with t^0 in all usage contexts that reflect one of its source senses (i.e., (c_i, y_i) where $y_i \in \mathcal{S}_0$), and replace w with t^* in all usage contexts where w signifies the target sense (i.e. (c_i, y_i) where $y_i = s^*$).

To guard against information smuggling in predicting novel word sense extension, we learn a contextualized language model from scratch using the set of replaced usage instances. Specifically, the language model is trained on the task of masked language modeling (MLM), where it takes batches of sampled usage instances with some randomly chosen tokens masked out, and updates its parameter weights to maximize the probability of infilling the correct missing tokens. Through this procedure, we obtain a language model that can compute meaningful contextualized representations for the usages of w that instantiate the target sense s^* *without* knowledge that s can be expressed by w.

3.2 Probabilistic formulation of WSE

Let $\mathcal{C}_0, \mathcal{C}_*$ be the two sets of usage instances with w replaced by t^* and t^0 respectively. We consider an inference scenario where the language model learned using the procedure from the previous section is presented with a novel usage $c^* \in \mathcal{C}_*$ of target token t^*, and is queried to choose among a set of candidate source tokens to convey the same (and new) intended meaning as that of t^*.

Concretely, suppose the target token t^* partitioned from the verb w = *arrive* denotes its metaphorical sense s^* = "to achieve a goal", and the source partitioned token t^0 of *arrive* is comprised of its existing source senses (that exclude the metaphorical sense in question). We then use the model to infer whether t^0 can be used to convey the new meaning t^* in novel metaphorical usages such as c = "They finally t^* at a conclusion after a long debate" (note here the original verb *arrive* is replaced by the target token t^* through word type partitioning). We assess the success of our model

by analyzing how it ranks the ground-truth source token (i.e., t^0 of *arrive*) among the space of alternative candidate source tokens partitioned from other polysemous words in the lexicon. For example, one source token might signify the literal senses of the verb *leave* which differs from the ground-truth verb *arrive*. Formally, we cast WSE as finding a source token t that maximizes the following probability:

$$\text{argmax}_t P(t|\mathbf{m}(t^*|c^*)) \quad (1)$$

Here $\mathbf{m}(t^*|c^*)$ is the representation of target token t^* under context c^* to which t is extended.

3.3 Chaining-based models of WSE

We present a family of probabilistic models for Eq.1 that draw inspirations from the cognitive theory of chaining (Lakoff, 1987; Habibi et al., 2020). Our chaining-based WSE models assume that a source token t^0 can be extended to express a novel meaning if the new intended meaning is overall similar to t^0's existing senses. We operationalize $\mathbf{m}(t^*|c^*)$ as the contextualized word embedding of target token t^* under context c^* computed by the speaker language model, denoted as $\mathbf{h}(t^*|c^*)$. We represent the existing senses of source token t as the collection of all of its contextualized embeddings $\mathbf{H}(t^0) = \{\mathbf{h}(t^0|c)|c \in \mathcal{C}_0\}$. The chaining-based WSE models take the general form:

$$P(t^0|\mathbf{m}(t^*|c^*)) \propto \text{sim}(\mathbf{H}(t^0), \mathbf{h}(t^*|c^*)) \quad (2)$$

We consider two common types of chaining model that specify the similarity function sim().

WSE-Prototype model. The prototype model takes inspiration from prototypical network for few-shot learning (Snell et al., 2017; Holla et al., 2020) and follows the prototype theory of categorization (Rosch, 1975) in cognitive psychology. It assumes that the existing senses of a source token t^0 can be summarized by a global average (i.e., prototype) of its contextualized embeddings in $\mathbf{H}(t^0)$, so that the probability of t^0 being a good candidate to convey the intended meaning of the target token is proportional to the semantic similarity between the contextualized embedding $\mathbf{h}(t^*|c^*)$ of the target token and the prototype of its sibling source token:

$$P(t^0|\mathbf{m}(t^*|c^*)) \propto \exp[-d(\mathbf{h}(t^*|c^*), \mathbf{z}(t^0))] \quad (3)$$

$$\mathbf{z}(t^0) = \frac{1}{|\mathcal{C}_0|} \sum_{c \in \mathcal{C}_0} \mathbf{h}(t^0|c) \quad (4)$$

Here $\mathbf{z}(t^0)$ is the global mean contextualized embedding of t^0, and we compute dot product as the similarity function $d(\cdot, \cdot)$ between two vectors.[2]

WSE-Exemplar model. The exemplar model resembles the memory-augmented matching network in deep few-shot learning (Vinyals et al., 2016), and formalizes the exemplar theory of categorization (Nosofsky, 1986). This model postulates that the meaning of t^0 is represented by the collection of its individual usages $c \in \mathcal{C}_0$. The probability that t^0 can be extended to the meaning $\mathbf{m}(t^*|c^*)$ is proportional to the mean similarity score between $\mathbf{h}(t^*|c^*)$ and each contextualized embedding of t^0:

$$P(t^0|\mathbf{m}(t^*|c^*)) \propto \frac{1}{|\mathcal{C}_0|} \sum_{c \in \mathcal{C}_0} \exp[-d(\mathbf{h}(t^*|c^*), \mathbf{h}(t^0|c))] \quad (5)$$

3.4 Learning sense-extensional semantic space

Chaining relies on identifying close semantic relations between existing senses and generalizing the recognized relations to generate new senses. For instance, if a WSE model has observed how the English verb *grasp* relates its literal sense "to hold an item firmly" to the extended metaphorical sense "to understand an idea", the model should also predict similar but novel non-literal sense extensions for other verbs that involve such metaphorical mappings (e.g., the meaning extension of the verb *get* from "to *get* a car" to "to *get* someone's idea", which also reflects the conceptual metaphor IDEAS ARE OBJECTS) (Lakoff and Johnson, 2008).

Following work in deep few-shot learning, we propose an episodic learning algorithm to transform the language model embedding space of the WSE model into a semantic space that better captures the regular, systematic patterns in sense extension. At each episode, we sample a mini-batch of N source-target token pairs $\{(t_i^0, t_i^*)\}_{i=1}^N$ partitioned from N distinct polysemous word types, and sample a usage context c_i^* for each target token t_i^*. The WSE model then chooses the most appropriate source token to convey the contextualized meaning of each target token. The parameter weights in the language model are optimized to minimize the negative log-likelihood of the ground-truth source

[2]We experimented with negative squared Euclidean distance suggested in Snell et al. (2017) as an alternative similarity function but found it to yield worse performance on both WSE and downstream WSD tasks compared to dot product.

token t_i^0 for each target token t_i^*:

$$\mathcal{J} = \sum_{i=1}^{N} -\log \frac{\text{sim}(\mathbf{H}(t_i^0), \mathbf{h}(t_i^*|c_i^*))}{\sum_{j=1}^{N} \text{sim}(\mathbf{H}(t_j^0), \mathbf{h}(t_i^*|c_i^*))} \quad (6)$$

Here $\text{sim}(\cdot,\cdot)$ can be either a prototype-based similarity function in Eq.3, or its exemplar-based counterpart specified in Eq.5.

4 Data

4.1 Dataset of polysemous word usages

We construct our WSE dataset by collecting naturalistic usage instances of English polysemous words from the Wikitext-103 linguistic corpus (Merity et al., 2016) that is commonly used as a language modeling benchmark. We first extract the sentences and lemmatize the corpus using SpaCy. We then apply a state-of-the-art word disambiguation algorithm by Bevilacqua and Navigli (2020) on each sentence to annotate each of its token with one of its associated WordNet synset IDs as the sense label (Miller, 1995). We construct a polysemous English word vocabulary by taking word lemma types that satisfy the following conditions: 1) the word type has least 2 different senses detected in the corpus; 2) each mention of the word type has one of the four part-of-speech categories as detected by SpaCy: noun, verb, adjective, or adverb; 3) each sense of the word type has at least 10 mentions in the corpus. This process yields a large repertoire of 7,599 polysemous word types with a total number of 1,470,211 usage sentences, and an average number of 4.27 senses per word type.

4.2 Partioning polysemous word types

To construct and evaluate our WSE framework, we partition each polysemous word types into multiple source-target pseudo-token pairs. In particular, for each word type w with n senses, we randomly choose one sense as the target sense s^*, and the remaining $n-1$ senses as the source senses. A source-target token pair is then created, which replace w in usage sentences based on their sense labels following the procedures described in Section 3.1. We repeat this partitioning process 5 times so that each word type with at least 5 senses will have 5 distinct senses chosen as target, and for words with less than 5 senses, the 5 target senses will be sampled with replacement from its sense inventory. Each partition will therefore create $2 \times 7,599 = 15,198$ pseudo-tokens.

5 Evaluation and results

5.1 Experimental setup

We use a transformer model with the same architecture as BERT-base-uncased (Devlin et al., 2019) as the main language model in our WSE framework. The parameter weights of our language models are randomly initialized to prevent any information smuggling (i.e., the models are trained from scratch). In the masked language modeling training stage on replaced usage sentences, we increase the vocabulary size of each model by replacing all polysemous word types in our WSE dataset vocabulary with their partitioned pseudo-tokens, and add rows to embedding layer and final classification layer of the BERT model accordingly. Five language models are trained independently, one for each set of partitioned tokens as described in section 4.2. During sense-extensional semantic space learning, we randomly choose 70% of the original polysemous word types and take usage sentences containing their partitioned tokens as the training set. Sentences containing partitioned tokens spawned by the remaining 30% word types will be taken as the test set, so that there is no overlap in the vocabulary of partitioned tokens or their parent word types between training and testing.[3]

5.2 Baseline models

We also compare the performance of our WSE models against a set of baseline models without chaining-based inference mechanisms: 1) a BERT-MLM baseline ignores the intended meaning information and predicts $P(t^0|\mathbf{m}(t^*|c^*))$ as the infilling probability of t^0 under context c^* with t^* replaced by a masking placeholder; 2) a BERT-STS baseline computes the contextualized representation $\mathbf{h}(t^0|c^*)$ of each candidate source token t^0 under c^*, and calculates $P(t^0|\mathbf{m}(t^*|c^*))$ as proportional to the cosine similarity between $\mathbf{h}(t^0|c^*)$ and the contextualized embedding $\mathbf{h}(t^*|c^*)$ of the target token under the same context (i.e. based on the semantic textual similarity between contextualized meanings of t^0 and t^*). Both baselines are built on the same BERT encoder just as the two chaining-based WSE models. We also consider a random baseline that randomly draws a source token from the set of alternative candidate tokens.

[3] See Appendix A for more implementation details.

Model	Mean reciprocal rank		Mean precision	
	Unsupervised	Supervised	Unsupervised	Supervised
Random Baseline	5.21	5.21	1.00	1.00
BERT-STS	11.89 (0.54)	33.55 (0.97)	14.02 (0.58)	25.57 (0.79)
BERT-MLM	15.57 (0.60)	37.09 (0.92)	16.34 (0.70)	28.99 (0.63)
WSE-Prototype	29.96 (0.77)	48.04 (1.03)	21.50 (0.44)	35.78 (1.16)
WSE-Exemplar	**34.25 (0.99)**	**53.79 (1.07)**	**29.17 (1.28)**	**37.82 (1.45)**

Table 1: Summary of model mean precision and MRR-100 scores (%) for word sense extension. Numbers after ± are standard deviations over 5 sets of independently partitioned source-target token pairs.

Model	Top-5 predicted words (source tokens)	Predicted rank of ground-truth source token
Word: *cover*; target sense: **be responsible for reporting news** Usage context: Generally, only reporters who *cover* breaking news are eligible.		
BERT-MLM	work, take, write, report, send	54/100
WSE-Exemplar	practice, report, supervise, cover, know	4/100
Word: *cell*; target sense: **a room where a prisoner is kept** Usage context: on the eve of his scheduled execution, he committed suicide in his *cell* with a smuggled blasting cap ...		
BERT-MLM	place, house, room, bedroom, hall	63/100
WSE-Exemplar	room, cell, bedroom, pocket, pyjamas	2/100
Word: *grasp*; target sense: **to get the meaning of** Usage context: Madonna later acknowledged that she had not *grasped* the concept of her mother dying.		
BERT-MLM	understand, remember, enjoy, comprehend, keep	82/100
WSE-Exemplar	understand, resolve, know, get, convey	43/100

Table 2: Example predictions made by the WSE-Exemplar model and the BERT-MLM baseline (supervised version). The top-5 predicted source tokens are translated into the (lemmatized) parent words from which they are partitioned.

5.3 Evaluation on WSE

We first evaluate our models on the task of predicting source partitioned tokens formulated in Eq.1. At each trial, for each target token t^*_w partitioned from w, we present the model with the ground-truth source token t^0_w partitioned from the same word w, and 99 negative candidate source tokens $t^0_{w'}$ spawned from different polysemous word types w'. Both the ground-truth source token and the negative candidates are sampled from the evaluation set for sense-extensional semantic space learning. We assess each model in two settings: an unsupervised version of a model that does not learn from the training set of WSE, and a supervised version that is trained on the training set of sense extensional space learning. The BERT encoders of the supervised versions of two BERT baselines are trained using the same objective function and data as defined in Section 3.4.

We quantify model performance with two metrics: 1) the mean precision is the percentage of cases where a model correctly predicts the ground-truth source token as the most likely candidate, and 2) the mean reciprocal rank (MRR-100) is the averaged multiplicative inverse of the ranks of the ground-truth source tokens in all evaluation examples. Table 1 summarizes the overall results in the five sets of independently partitioned tokens. We make several observations: 1) all BERT-based models perform substantially better than chance even without explicit training on WSE. This can be explained by the fact that many polysemous word types in our dataset have very fine-grained WordNet senses, so that the target senses chosen from its sense inventory are often highly similar or even hardly distinguishable from the some source senses of the same word; 2) all BERT-based models benefit from learning a sense-extensional semantic space, suggesting the presence of regularity shared among examples of sense extension across word types; 3) both chaining-based WSE models consistently outperform other baselines in both the unsupervised and supervised settings. The exemplar-based WSE models generally outperform than their prototype-based counterparts, suggesting that word sense extension depends on the speaker's sensitivity to the semantic similarity between the intended meaning and the individual (exemplar) usages.

Table 2 shows example predictions on sam-

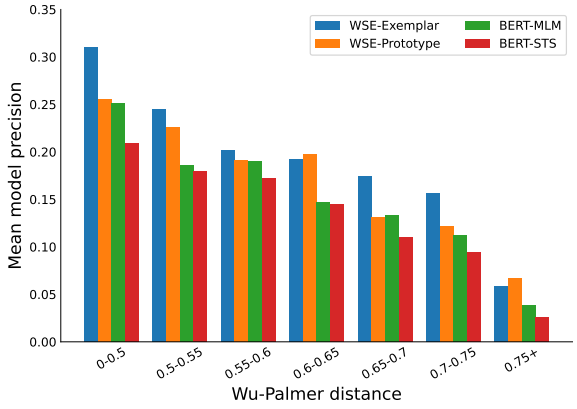

Figure 2: Mean model precision vs. Wu-Palmer distance between WordNet synsets associated with fully-partitioned tokens.

ple polysemous words made by the supervised exemplar-based WSE model and the supervised BERT-MLM baseline. The WSE model successfully predicts many types of sense extension, such as metaphorical senses for both the verb *cover* example and the noun *cell*. In contrast, the BERT-MLM baseline shows a greater tendency to predict a literal paraphrase for a partitioned token. Still, both WSE and baseline models struggle with predicting some usages that involve strong non-literal sense extension (e.g., the *grasp* example).

5.4 Sense relatedness and model predictability

Prior work in psycholinguistics suggests that both adults and children often find it easy to infer a new intended meaning of a word if they can access a highly related conventional sense of that word to constrain their interpretation (Clark and Gerrig, 1983; Klepousniotou et al., 2008; Rodd et al., 2012). We examine whether our WSE models exhibit human-like sensitivity to the conceptual relatedness between existing and novel word senses. For each source-target partitioned token pair (t^0, t^*), we quantify their degree of conceptual relatedness as the mean Wu-Palmer semantic distance (Wu and Palmer, 1994) between the WordNet synset of the target sense denoted by t^* and the synset of each existing source sense of t^0. Figure 2 shows the performance of 4 WSE model variants on predicting sense pairs binned with respect to their degree of conceptual similarity. We observe that the WSE models generally make better predictions on source-target token pairs that are semantically more related (e.g., metonymy), and perform less well on examples where the target sense is conceptually very different to the existing source senses (e.g., strong metaphor or homonymy).

5.5 Application of WSE to WSD

As a final step, we show that state-of-the-art word sense disambiguation models can benefit from the word sense extension framework. We evaluate WSD models on the standard WSD evaluation framework proposed by (Raganato et al., 2017), where in each trial, the model is given an input sentence and is asked to assign WordNet sense labels for a subset of tokens within the sentence. We consider two BERT-based WSD models: 1) a BERT-linear model that learns a linear classifier for WSD on top of a frozen BERT encoder. This model does not incorporate gloss information, and cannot predict novel senses that do not appear in training; 2) a bi-encoder model (BEM) by (Blevins and Zettlemoyer, 2020) independently encodes input sentences with target words and sense glosses via two encoders, each of which are initialized with BERT-base. The contextualized embedding of the target word then takes dot product with the gloss embedding of each candidate sense, and the model predicts the sense with highest dot product score with the embedded target word. This model has been shown to yield impressive results on WSD examples with rare senses.

To integrate WSE into WSD, we fine-tune the BERT encoder of each WSD model on the WSE training set of Wikitext-103 usage sentences via the objective in Eq. 6, which can be formulated as either a prototype model or an exemplar model. Unlike the case of WSE evaluation, here we use pretrained BERT-base-uncased encoders and keep the original word form of each polysemous word without partitioning it into source-target token pairs. The resulting BERT encoder is then taken to learn one of the two WSD models described above, and evaluated on WSD tasks. For BEM, both encoders are initialized as the BERT-base fine-tuned on WSE. Since the sense labels of usage sentences in the WSE dataset are not fed to BERT during training, none of the models has access to any usage examples of target senses in the WSD test set.

Table 3 reports overall results on the WSD datasets under the standard F1-score. We also include the performance of two simple baselines: 1) WordNet S1 always predicts the first sense, and 2) MFS always predicts the most frequent sense in the training data. We found that chaining-based WSE

	Dev	Test Datasets				Concatenation of Test Datasets				
	SE07	SE02	SE03	SE13	SE15	Nouns	Verbs	Adj.	Adv.	ALL
WordNet S1	55.2	66.8	66.2	63.0	67.8	67.6	50.3	74.3	80.9	65.2
Most frequent sense (MFS)	54.5	65.6	66.0	63.8	67.1	67.7	49.8	73.1	80.5	65.5
BERT-linear	68.6	75.2	74.7	70.6	75.2	74.6	63.6	78.6	87.0	73.5
+ WSE-Prototype	**70.9**	**78.0**	**75.2**	**71.2**	**77.9**	**75.5**	**66.1**	78.9	**87.1**	**76.4**
+ WSE-Exemplar	70.5	**78.0**	75.1	**71.2**	77.7	74.8	65.8	**79.2**	86.4	75.3
BEM	74.3	78.8	**77.4**	79.6	80.9	81.5	68.5	82.8	**87.1**	78.8
+ WSE-Prototype	**74.9**	**80.2**	75.9	**81.2**	81.1	**82.5**	**70.2**	**83.9**	**87.1**	**80.1**
+ WSE-Exemplar	74.5	80.0	76.1	**81.2**	**81.7**	81.4	69.1	81.2	86.4	79.2

Table 3: F1-scores (%) for fine-grained all-words WSD task on the evaluation framework by (Raganato et al., 2017).

WSD test example	BEM prediction (no WSE)	BEM prediction (with WSE)
Context: The purpose of education is to encourage young men and women to realize their ***full*** academic potential. **Target sense training frequency**: 0	containing as much or as many as is possible (✗)	complete in extent or degree (✓)
Context: Haney felt like ***shrinking*** out of sight, but he was already trapped in the corner with the wiry, dark little man. **Target sense training frequency**: 1	reduce in size/physically (✗)	draw back with fear or pain (✓)

Table 4: Examples of context and definitions of WSD-model predicted senses. The bold italic words in context are disambiguated by the BEM model before and after training on WSE.

	Sense frequency		
	High	Few-shot	Zero-shot
BERT-linear	81.7	54.4	53.6
+ WSE	82.3	60.1	53.6
BEM	86.8	77.7	67.8
+ WSE	86.6	79.6	71.5

Table 5: F1-score (%) on subsets of the WSD test dataset grouped by target sense frequency in SemCor corpus.

models improve the performance of the two BERT-based WSD models on almost every test subset, as well as on all POS categories except for the adverb class. These results show that WSE may serve as useful pretraining for improving WSD models both with and without access to gloss information.

Rare word-sense pairs. We hypothesize that WSE improves WSD because learning word sense extension helps the model to better interpret rare senses that bear systematic semantic relations with more conventional senses. Table 5 shows the performance of WSD models grouped by the frequency of the target word sense in the WSD training set. We define zero-shot test cases as target senses that never appear during WSD training, and few-shot test cases as those with 1 to 10 mentions, and high-frequency senses as those with more than 10 training mentions. The BERT-linear model resort to a most frequent sense heuristic for zero-shot examples, since it cannot learn a classification layer embedding for previously unattested senses. We observe that all WSD models trained on WSE yield substantially greater improvement for few-shot and zero-shot test cases, while maintaining high performance on the more frequent cases. Table 4 shows test examples where incorrect predictions of BEM are improved with WSE integration. These examples often exhibit regular semantic relations between target and conventional senses of a word (e.g., the relation between physical size and amount that underlies the two attested senses of *full*).

6 Conclusion

We have presented a framework for word sense extension that supports lexical items to extend to new senses in novel context. Our results show that chaining provides a general mechanism for automated novel sense extension in natural context, and learning a transformed sense-extensional space enables systematic generalization to a certain degree. We also show that word sense extension improves

the performance of transformer-based WSD models particularly on rare word senses. Future work may extend our framework in several ways, such as how to better model systematic word sense extension, and do so over time and in different languages.

7 Ethical considerations

We discuss the limitations and potential risks of our work.

7.1 Limitations

Our current framework does not explicitly consider the temporal order via which word senses have emerged. In particular, in the data collection step, we construct source-target token pairs for each word type by randomly sampling a target sense from its sense inventory. An alternative and more realistic approach would be to sort all senses of a word chronologically by their times of emergence in history, and use the model to incrementally predict each sense of a word based on usages of its older senses. However, we found that it is infeasible to find accurate timestamps of senses in natural corpora at a comprehensive scale. Another approach is to have human annotators evaluate the plausibility of each ground-truth source-target token pairs that are automatically created in our data collection pipeline, which is a potential area for future consideration.

7.2 Potential risks

All scientific artifacts in this study have been made publicly available and are consistent with their intended use and access conditions. We acknowledge that our focus on English might introduce linguistically or culturally specific biases in model-generated outputs. For instance, we observe that the WSE models trained on English sentences learn to generate a metaphorical expression "to *spend* some time" for the English verb *spend*, which is common in English but differ in other languages (e.g., Hungarian speakers instead tend to say "to *fill* some time" as in Kövecses et al. 2010). We believe that by training WSE models cross-linguistically to cover various innovative lexical uses should help mitigate this issue.

8 Acknowledgements

This work was supported by a NSERC Discovery Grant RGPIN-2018-05872.

References

Michele Bevilacqua and Roberto Navigli. 2020. Breaking through the 80% glass ceiling: Raising the state of the art in word sense disambiguation by incorporating knowledge graph information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2854–2864.

Terra Blevins, Mandar Joshi, and Luke Zettlemoyer. 2021. Fews: Large-scale, low-shot word sense disambiguation with the dictionary. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 455–465.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

Claudia Brugman and George Lakoff. 1988. Cognitive topology and lexical networks. In *Lexical ambiguity resolution*, pages 477–508. Elsevier.

Howard Chen, Mengzhou Xia, and Danqi Chen. 2021. Non-parametric few-shot learning for word sense disambiguation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1774–1781, Online. Association for Computational Linguistics.

Herbert H Clark and Richard J Gerrig. 1983. Understanding old words with new meanings. *Journal of verbal learning and verbal behavior*, 22(5):591–608.

Ann Copestake and Ted Briscoe. 1995. Semi-productive polysemy and sense extension. *Journal of semantics*, 12(1):15–67.

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.

Robert B Dewell. 1994. Overagain: Image-schema transformations in semantic analysis. *Cognitive Linguistics*, 5(4).

Haim Dubossarsky, Eitan Grossman, and Daphna Weinshall. 2018. Coming to your senses: on controls and evaluation sets in polysemy research. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1732–1740, Brussels, Belgium. Association for Computational Linguistics.

Brian Falkenhainer, Kenneth D Forbus, and Dedre Gentner. 1989. The structure-mapping engine: Algorithm and examples. *Artificial intelligence*, 41(1):1–63.

William A Gale, Kenneth Church, and David Yarowsky. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. In *30th Annual Meeting of the Association for Computational Linguistics*, pages 249–256.

Aina Garí Soler and Marianna Apidianaki. 2021. Let's play mono-poly: Bert can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Dedre Gentner. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive science*, 7(2):155–170.

Raymond W Gibbs Jr and Herbert L Colston. 2008. Image schema. In *Cognitive Linguistics: Basic Readings*, pages 239–268. De Gruyter Mouton.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3960–3973, Online. Association for Computational Linguistics.

Karan Grewal and Yang Xu. 2020. Chaining and historical adjective extension. In *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*.

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

Christian Hadiwinoto, Hwee Tou Ng, and Wee Chung Gan. 2019. Improved word sense disambiguation using pre-trained contextualized word representations. 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 5297–5306.

Nithin Holla, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Learning to learn to disambiguate: Meta-learning for few-shot word sense disambiguation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4517–4533, Online. Association for Computational Linguistics.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuan-Jing Huang. 2019. Glossbert: BERT for word sense disambiguation with gloss knowledge. 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 3509–3514.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907.

Adam Kilgarriff. 2004. How dominant is the commonest sense of a word? In *Text, Speech and Dialogue: 7th International Conference, TSD 2004, Brno, Czech Republic, September 8-11, 2004, Proceedings*, volume 3206, page 103. Springer Science & Business Media.

Adam Kilgarriff and Joseph Rosenzweig. 2000. Framework and results for english senseval. *Computers and the Humanities*, 34(1):15–48.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Ekaterini Klepousniotou, Debra Titone, and Carolina Romero. 2008. Making sense of word senses: the comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 34(6):1534.

Zoltán Kövecses et al. 2010. Metaphor and culture. *Acta Universitatis Sapientiae, Philologica*, 2(2):197–220.

Sawan Kumar, Sharmistha Jat, Karan Saxena, and Partha Talukdar. 2019. Zero-shot word sense disambiguation using sense definition embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5670–5681.

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

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Adrienne Lehrer. 1990. Polysemy, conventionality, and the structure of the lexicon. *Cognitive Linguistics*, 1(2).

Daniel Loureiro and Alipio Jorge. 2019. Language modelling makes sense: Propagating representations through wordnet for full-coverage word sense disambiguation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691.

Fuli Luo, Tianyu Liu, Qiaolin Xia, Baobao Chang, and Zhifang Sui. 2018. Incorporating glosses into neural word sense disambiguation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2473–2482.

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(2):230–262.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

Oren Melamud, Jacob Goldberger, and Ido Dagan. 2016. context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 51–61, Berlin, Germany. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Robert M Nosofsky. 1986. Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39.

Geoffrey Nunberg. 1979. The non-uniqueness of semantic solutions: Polysemy. *Linguistics and philosophy*, pages 143–184.

Mohammad Taher Pilehvar and Roberto Navigli. 2014. A large-scale pseudoword-based evaluation framework for state-of-the-art word sense disambiguation. *Computational Linguistics*, 40(4):837–881.

James Pustejovsky. 1998. *The generative lexicon*. MIT press.

James Pustejovsky and Anna Rumshisky. 2010. Mechanisms of sense extension in verbs.

Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110.

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(10):2323–2328.

Jennifer M Rodd, Richard Berriman, Matt Landau, Theresa Lee, Carol Ho, M Gareth Gaskell, and Matthew H Davis. 2012. Learning new meanings for old words: Effects of semantic relatedness. *Memory & Cognition*, 40(7):1095–1108.

Eleanor Rosch. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104(3):192.

Anna Rumshisky and Olga Batiukova. 2008. Polysemy in verbs: Systematic relations between senses and their effect on annotation. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 33–41, Manchester, UK. Coling 2008 Organizing Committee.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087.

Zhewei Sun, Richard Zemel, and Yang Xu. 2021. A computational framework for slang generation. *Transactions of the Association for Computational Linguistics*, 9:462–478.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in Neural Information Processing Systems*, 29:3630–3638.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Gregor Wiedemann, Steffen Remus, Avi Chawla, and Chris Biemann. 2019. Does bert make any sense? interpretable word sense disambiguation with contextualized embeddings. *arXiv preprint arXiv:1909.10430*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138.

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.

Zhi Zhong and Hwee Tou Ng. 2010. It makes sense: A wide-coverage word sense disambiguation system for free text. In *Proceedings of the ACL 2010 system demonstrations*, pages 78–83.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. BERT-based lexical substitution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3368–3373, Florence, Italy. Association for Computational Linguistics.

A Implementations of WSE models

We use the BERT-base-uncased configuration provided by Hugging Face (Wolf et al., 2020) to initialize all BERT-based WSE models (two baselines and two chaining-based models). During MLM pretraining of BERT models on replaced usage sentences by partitioned pseudo-tokens, we randomly mask 15% of tokens in each sentence, and train each model on predicting the masked tokens. We add all partitioned pseudo-tokens as special tokens into the vocabulary of the BERT tokenizer, so each pseudo-token will be encoded as a whole in the input sequence. Learning is performed using the Adam optimizer (Kingma and Ba, 2015), with a learning rate of 5e-5 and a batch size of 128, for 8 epochs (after which all models achieved highest evaluation accuracy). During sense-extensional semantic space learning, both exemplar-based and prototype-based models are trained on the objective function in Eq.6 using Adam, with a mini-batch size of 16 and a learning rate of 2e-5, for 8 epochs (after which all models achieved highest evaluation accuracy). All experiments are run on machines with 4 NVIDIA Tesla V100 GPUs, with an average training time of 30 minutes per epoch for MLM pretraining, and 12 minutes per epoch for sense-extensional semantic space learning.

ACL 2023 Responsible NLP Checklist

A For every submission:

☐ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Left blank.

☐ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☐ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☐ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☐ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Left blank.

D ☐ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.