Word sense generation

Anonymous ACL submission

Abstract

The lexicon makes creative reuse of words to express novel senses. A long-standing effort in natural language processing has been focusing on disambiguating and inducing word senses from context. Little has been explored about how novel word senses may be generated automatically. We consider a paradigm of *word sense generation* (WSG) that enables words to spawn new senses by extending toward novel naturalistic context. We develop a general framework that simulates novel word sense extension by dividing a word into hypothetical child tokens and making inferences about the plausibility of sense extension among the sibling tokens in usage sentences that never appear in training. Our framework combines probabilistic models of chaining with a learning scheme that transforms a language model embedding space to support various types of word sense extensions. We evaluate our framework rigorously against several competitive baselines and show that it is superior in predicting plausible novel senses including metonymic and metaphoric word usages in a large set of 1,500 English verbs. We show that the learned semantic space exhibits systematic patterns of word sense extension while retaining competence in common natural language processing tasks.

1 Introduction

A key property of the lexicon is the creative reuse of words to express novel senses. For example, the English phrase *to arrive at* extended from its original sense "to reach a physical location (e.g., gate)" toward new senses such as "to come to an event (e.g., concert)" and "to reach an abstract, cognitive state (e.g., conclusion)". The extension of word meaning toward new context may appear to draw on different processes ranging from metonymy to metaphor (see Figure 1 for an illustration), but here we present a general framework that infers how words extend to plausible new senses under novel naturalistic context.

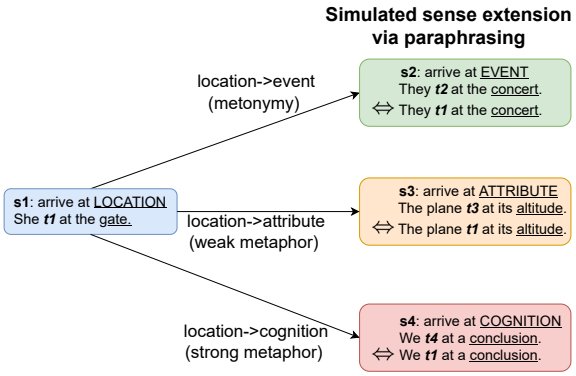


Figure 1: The word sense generation (WSG) framework. A verb (e.g., *arrive*) with a set of sense-labeled usage sentences is partitioned into distinct child tokens (e.g., t_1-t_4 signifying senses s_1-s_4 as illustrated). Each token represents a hypothetical word type that replaces its original parent verb in sentences where it expresses a given sense (e.g., *arrive* in sentences where it predicates a location will be substituted by the hypothetical t_1, and separately substituted by other tokens t_2-t_4 for sentences expressing the other senses). WSG models infer whether a child token (e.g., t_1) can be extended to express the senses of its siblings via paraphrasing novel sentences that do not appear in training (e.g., t_1 has 3 potential siblings to extend its meaning to: t_2, t_3, t_4).

One of the most long-standing efforts in natural language processing (NLP) is to be able to disambiguate word senses from text. This line of work takes a discriminative approach toward the multifaceted aspect of word meaning and has developed models relying on both traditional machine learning techniques (Gale et al., 1992; Kilgarriff and Rosenzweig, 2000; Zhong and Ng, 2010; Iacobacci et al., 2016; Raganato et al., 2017) and modern neural language models (Huang et al., 2019; Wiedemann et al., 2019; Loureiro and Jorge, 2019; Bevilacqua and Navigli, 2020). Related work has developed automated methods for inducing or detecting novel word senses (Lau et al., 2012; Cook et al., 2014; Lau et al., 2014), which can also be considered as unsupervised approaches to sense

disambiguation. Here we consider an alternative paradigm that aims to model word senses by taking a generative approach in naturalistic context.

Work in computational and cognitive linguistics suggests that word senses often do not extend arbitrarily (Nunberg, 1979; Lehrer, 1990). Lexical semanticists have pointed out that a number of cognitive devices may be applied to generate creative word uses, 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 meanings (Brugman and Lakoff, 1988; Lakoff and Johnson, 2008). However, the reliance on hand-crafted rules of semantic productivity makes it difficult to implement these accounts into computational systems that support flexible and scalable generation of novel word senses.

We develop a principled framework termed *word sense generation* (WSG). As a starting point, we focus on modelling sense generation of English verbs, which constitute a broad yet notoriously challenging class of productive sense extensions (Pustejovsky and Rumshisky, 2010). Figure 1 illustrates our WSG framework. Given a verb (e.g., *arrive*), we consider a paradigm that simulates how it can be extended to express plausible novel senses in naturalistic context. We do so by dividing a word into hypothetical child tokens signifying its different senses. We then infer whether a child token may be extended to express its sibling tokens under novel context (i.e., a simulation of how a word might extend from its existing senses to novel senses). We propose a family of deep probabilistic models for this inference problem that are built on the cognitive theory of chaining, which states that word meanings grow by linking novel senses to existing ones that are close in semantic space (Lakoff, 1987; Malt et al., 1999; Ramiro et al., 2018; Habibi et al., 2020; Yu and Xu, 2021). We expect these chaining models to support the incremental extension of a word's meaning toward a variety of new senses, a process analogous to the gradient cline of verb sense extensions (e.g., weak metaphor) discussed in Pustejovsky and Rumshisky (2010).

We make three contributions: 1) we formulate word sense generation as a novel probabilistic inference task whereby a language model, after learning a set of partitioned tokens signifying different senses of a polysemous word, automatically infers whether sibling tokens can be used interchangeably under novel context absent in training; 2) we develop a family of WSG models motivated by the cognitive theories and models of semantic chaining, and a new learning scheme to capture regular patterns of word sense extension; 3) we collect a new dataset of word sense generation examples which includes natural usages for approximately 22,000 senses of over 1,500 common English verbs.[1]

2 Related work

2.1 Theories 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 and Johnson (2008) argues that systematic mappings across conceptual domains result in the abundance of metaphorical word senses in natural language, while these mappings can be motivated by cognitive processes such as chaining (Lakoff, 1987) and image schema transformation (Brugman and Lakoff, 1988; Dewell, 1994; Gibbs Jr and Colston, 2008). Our work connects the cognitive and formal approaches to word meaning extension by showing that a cognitively inspired chaining-based word sense generation framework can learn systematic patterns of meaning extension discussed in the tradition of generative lexical semantics.

2.2 Non-literal language generation

Our framework also relates to research on automated generation of non-literal word uses such as metaphor and metonymy. Recent work has explored using contextualized language models to generate metaphorical paraphrases for literal usage sentences of a word (Tong et al., 2021; Stowe et al., 2021; Chakrabarty et al., 2021). It has also been shown that chaining mechanisms can be incorporated into contextualized language models to pre-

[1] We release the code and data for our work here: `PLACEHOLDER link`.

dict unconventional word usages such as slang (Sun et al., 2021). On the other hand, generation of more conventionalized senses, such as metonymy, have remained underexplored (Rambelli et al., 2020), because most contextualized language models have already been exposed to usages of the most common senses of a word type during pretraining, while evaluation on sense generation requires models without such prior knowledge. Our framework circumvents this circularity problem by creating novel tokens that reflect partial usages of a polysemous word and utilizing language models to learn them from scratch, thereby allowing us to model the generative processes of word senses in a zero-shot setting.

3 Framework of word sense generation

Our framework of word sense generation involves three interrelated components: 1) A procedure for partitioning polysemous words in the lexicon into child tokens signifying their different senses; 2) a probabilistic formulation for inferring a child token to paraphrase one of its siblings under a novel linguistic context; and 3) a representational learning algorithm for a transformed semantic space to learn flexible extensions of word senses.

3.1 Sense-based word type partitioning

Let $V = \{w_1, ..., w_{|V|}\}$ be our vocabulary of polysemous English verbs, where each verb w_i has a sense inventory $S(w_i) = \{s_i^{(1)}, ..., s_i^{(n_i)}\}$. Assume that for each sense s_i^j of w_i, there is a collection of its representative usage sentences $U(s_i^j)$ in which w_i exhibits sense s_i^j. We wish to investigate whether a language model, which has knowledge only about a partial set of senses of w_i, is able to generate a usage that reflects a novel sense of w_i. In particular, we define a *partition* of a word type w as a grouping of its sense inventory $S(w_i)$ into a collection of K_i distinct sense subsets $\{S_1^i, ..., S_{K_i}^i\}$ – for example, as illustrated in Figure 1, the sense inventory of the verb *arrive* $S = \{s_1, s_2, s_3, s_4\}$ can be partitioned into four singleton sets (represented by the four colored rectangles). For each sense subset S_k^i with an associated usage sentence set $U(S_k^i) = \bigcup_{s \in S_k^i} U(s)$, we replace all mentions of w_i in each $u \in U(S_k^i)$ with a novel child token t_k^i (e.g., the sentence "We *arrive* at a conclusion" in which *arrive* expresses the abstract state achievement sense s_4 will be converted into "We t_4 at a conclusion"). We then use a contextualized language model to learn semantic representations for each child token from the replaced usage sentences $U(S_k^i)$ via the task of masked language modeling (MLM). To prevent information smuggling, the language model is initialized from scratch and therefore does not have any *a priori* knowledge about either the partitioned tokens or their parent verb types. Next, we explain how the task of WSG can be formulated as a paraphrasing problem of inferring a partitioned child token to substitute one of its siblings under a given context.

3.2 Probabilistic formulation of WSG

3.2.1 WSG as partitioned token paraphrasing

Let S_k, S_l be two partitioned sense subsets of the sense inventory $S(w)$ of w, and t_k, t_l be their corresponding child tokens. We say that a language model generates a novel sense for token t_k (called the *source* token) if infers that t_k can serve as a good paraphrase token to substitute its sibling t_l (called the *target* token) in a usage sentence $u^* \in U(S_l)$ containing t_l that does not appear in the MLM training set. For instance, if the LM initially learns two child tokens spawned from the the verb *arrive* that reflect its two distinct senses s_1 = "to come to a physical location" and s_2 = "to achieve a goal" respectively, we would expect the LM to predict that the source token t_1 denoting the concrete sense s_1 can be used to paraphrase its target sibling t_2 denoting the abstract sense s_2 in usages such as "They t_2 at a conclusion after a debate". We cast WSG as inference of the following word choice probability:

$$P(t_k \rightarrow S_l; u^*) = P(t_k \Leftrightarrow t_l | u^*) \quad (1)$$

Here $t_k \rightarrow S_l$ means that t_k can be extended to express novel senses drawn from S_l, and $t_k \Leftrightarrow t_l$ means that t_k, t_l can be used interchageably under context u^*. Next, we introduce several models that infer the paraphrase probability in Eq.1.

3.2.2 Baseline models of WSG

We first consider two simple baseline models: the masked language modeling (MLM) baseline ignores information about t_l and predicts $P(t_k \Leftrightarrow t_l | u^*)$ simply as the infilling probability of t_k under a masked sequence of u^*:

$$P(t_k \Leftrightarrow t_l | u^*) = P_{MLM}(t_k | u^*_{\setminus t_l}) \quad (2)$$

Here $P_{MLM}(t_k|u^*_{\setminus t_l})$ denotes the probability of choosing t_k to infill a masked sequence of u^* with t_l replaced by a placeholder, as determined by the contextualized language model. The semantic textual similarity (STS) baseline instead predicts the paraphrase probability as proportional to the cosine similarity between the contextualized representations $h(t_l|u^*), h(t_k|u^*)$ of t_l and t_k under the context u^*:

$$P(t_k \Leftrightarrow t_l|u^*) \propto \text{cosine-sim}(h(t_l|u^*), h(t_k|u^*)) \quad (3)$$

3.2.3 Chaining-based models of WSG

A common issue of the two baseline models described is that they do not model the relations or semantic similarities between the the novel use of t_k in u^* and its existing usages $U(S_k)$. We therefore propose a family of WSG models that draws inspirations from the cognitive theory of chaining and memory-augmented deep learning. These models predict that a token t_k is a good paraphrase for its sibling t_l under u^* if the collection of conventional usages $U(S_k)$ of t_k bears a close overall proximity to u^* in the contextualized embedding space:

$$P(t_k \Leftrightarrow t_l|u^*) \propto \text{sim}(t_k, t_l) \quad (4)$$
$$= \text{sim}(H(t_k), h(t_l; u^*)) \quad (5)$$

Here $H(t_k) = \{h(t_k; u)\}_{u \in U(S_k)}$ is the collection of contextualized embeddings of t_k in its conventional usages. We next describe two commonly used chaining models that specify the similarity function $\text{sim}(H(t_k), h(t_l; u^*))$.

WSG-Prototype model. The prototype model draws inspirations from prototypical network for few-shot learning (Snell et al., 2017) and follows the prototype theory of categorization (Rosch, 1975) in cognitive psychology. It assumes that the meaning of a child token t_k can be summarized by a global mean of its contextualized embeddings taken from all of its conventional usages $U(S^i_k)$, so that the probability of t_k being a good paraphrase for t_l under u^* is proportional to the semantic similarity between the contextualized embedding $h(t_l|u^*)$ of t_l and the summary prototype of its sibling:

$$\text{sim}(t_k, t_l) = \exp(-\|h(t_l|u^*) - z(t_k)\|^2) \quad (6)$$
$$z(t_k) = \frac{1}{|U(S^i_k)|} \sum_{u \in U(S^i_k)} h(t_k|u) \quad (7)$$

Here $z(t_k)$ is the mean contextualized embedding of t_l over all of its existing usages, and we have defined semantic similarity as the negative exponential Euclidean distance between two embeddings.

WSG-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_k is represented by the entire collection of its usages $u \in U(S^i_k)$. The probability that t_k paraphrases t_l under context u^* is then proportional to the mean negative exponential Euclidean distance between $h(t_l|u^*)$ and each contextualized embedding of t_k:

$$\text{sim}(t_k, t_l) = \frac{1}{|U(S_k)|} \sum_{u \in U(S_k)} \exp(-\|h(t_l|u^*) - h(t_k|u^*)\|^2) \quad (8)$$

3.3 Learning sense-extensional semantic space

Chaining relies on identifying close semantic relations between senses, and we therefore develop a learning scheme that transforms a standard semantic space to one that is sensitive to regular relations attested in sense extension. For instance, if a WSG model has observed how verb *grasp* relates its literal sense (e.g., to *grasp* an item) to the extended metaphorical sense (e.g., to *grasp* an idea), under the transformed semantic space the model should also predict similar but novel non-literal sense extensions for other verbs that involve such metaphorical mappings (e.g., to *get* someone's idea, which also reflects the conceptual metaphor IDEAS ARE OBJECTS).

We follow work in deep few-shot learning and propose an episodic learning algorithm for a sense-extensional semantic space: at each episode, we sample a mini-batch of N source-target token pairs $\{(t^{(n)}_k, t^{(n)}_l)\}^N_{n=1}$ partitioned from N different parent word types, and sample a usage sentence $u^{*(n)}$ for each target token $t^{(n)}_l$. The model then learns to perform in-batch WSG by choosing a paraphrase token for every target $t^{(n)}_l$ in $u^{*(n)}$ among the set of N candidate source tokens $\{(t^{(1)}_k, ...t^{(N)}_k\}$, with $t^{(n)}_k$ being the ground-truth paraphrase. For the two chaining-based models, learning can be performed by minimizing the in-batch classification loss:

$$\mathcal{J} = \sum_{n=1}^{N} -\log \frac{\text{sim}(H(t^{(n)}_k), h(t^{(n)}_l; u^{*(n)}))}{\sum_{n'} \text{sim}(H(t^{(n')}_k), h(t^{(n)}_l; u^{*(n)}))} \quad (9)$$

Sense type	No. of verb types	No. of spawned tokens (Full partition)	No. of spawned tokens (Leave-one-out partition)	No. of usage sentences
Domain-based	1,199	14,735	5,860	481,654
Synset-based	1,468	6,463	2,936	371,920

Table 1: Summary statistics of the two collected WSG datasets.

Here $\text{sim}(\cdot,\cdot)$ can either be a prototype-based similarity function in Eq.6, or be its exemplar-based counterpart specified in Eq.8. The two baseline models can also be trained directly on the same set of examples by maximizing their word choice probabilities (i.e., Eq.2 and Eq.3).

4 Data

We collect usage sentences for English verbs from the Wikitext-103 linguistic corpus (Merity et al., 2016) that is commonly used as a language modeling benchmark dataset. It is well-known that word senses are highly fuzzy linguistic categories, and there does not exist a set of word senses that is suitable for every NLP task (Kilgarriff, 1997; Rumshisky and Batiukova, 2008). We therefore consider two different definitions of verb senses in our datasets, as well as two different ways of partitioning the sense inventory of a verb.

4.1 Domain-based vs. synset-based verb sense

Our first sense definition of an English verb is based on the semantic domain of its syntactic object nouns. In particular, for each usage of a verb with a noun object, we label the noun with its supersense category defined in the WordNet lexical database (Miller, 1995), and define a verb sense as the collection of usages whose object nouns share the same supersense label. This sense definition helps us investigate the regularity underlying sense extension cases of different words. For instance, a common type of logical polysemy involves coercing an event nominal to denote the physical location where it takes place (e.g. to arrive at the *theatre* $\rightarrow$ to arrive at the *concert*), which can be captured by extending the domain-based sense "to arrive at a LOCATION" to "to arrive at an EVENT" [2]. We also consider a second, more established type of word sense as recorded in lexicographic resources. In particular, for each word, we apply a state-of-the-art word sense disambiguation algorithm (Bevilacqua and Navigli, 2020) on each of its usage sentence to identify the its evoked WordNet sense, as indicated by one of its associated synset ID.

4.2 Full vs. leave-one-out word type partition

For each WSG dataset, we also consider two types of sense inventory partitioning: the *full partition* creates a child token for each sense of a parent verb (e.g., the partition in Fig.1), and a WSG model is evaluated on predicting extensions between all possible child token pairs spawned from the same word type. The *leave-one-out partition* instead randomly samples one sense for each verb to create the target token, and group all remaining senses into a single source token. Summary statistics of the two resulting datasets are shown in Table 1.

5 Evaluation and results

5.1 Model implementation

We use a BERT language model (Devlin et al., 2019) to build both the contextualized baselines and the chaining-based WSG models. All BERT models are implemented from scratch and are not pretrained on any NLP tasks. In the masked language modeling step, we increase the vocabulary size of each model by replacing all parent verbs in our datasets with their spawned child tokens, and increase the size of the model's embedding layer and final classification layer accordingly. During sense-extensional semantic space learning, we randomly choose 70% of the parent verb types in each dataset, and take usage sentences containing their spawned tokens as training set. Sentences containing partitioned tokens spawned by the remaining 30% verb types will be taken as the test set, so that there is no overlap in the vocabulary of partitioned tokens or parent verb types between training and testing.[3]

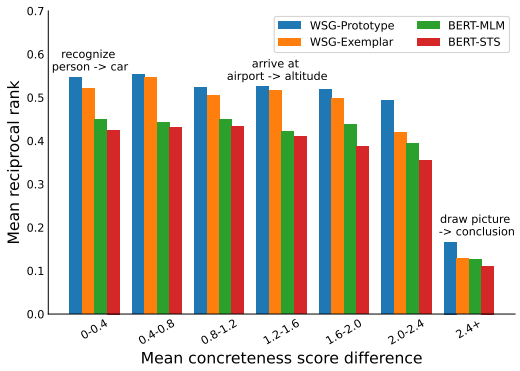
5.2 Evaluation on WSG

We first evaluated our models on the task of predicting paraphrase partitioned tokens formulated in Eq.1: given a target token and its sample usage

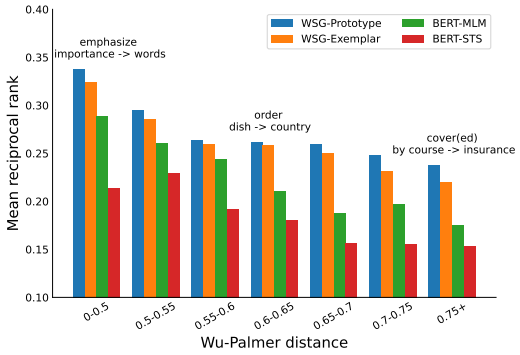
[2] See Appendix B for a full list of supersense categories and their descriptions.

[3] Implementation details are described in Appendix A.

sentence, how likely is the model to predict one of its sibling source tokens as a good paraphrase? At each trial, we pick one ground-truth source token and 99 negative tokens with different parent verb types, and ask the model to rank the 100 candidates based on their infilling likelihoods. We then compute the mean reciprocal ranks for each ground-truth source token among the 100 candidates (MRR-100) to assess model performance.



(a) Model preformance (MRR) vs. Concreteness score difference between supersenses.



(b) Model preformance (MRR) vs. Wu-Palmer distance between WordNet synsets.

Figure 2: Model performance vs. sense relatedness between fully-partitioned tokens.

Table 2 summarizes the averaged results over five independently sampled candidate sets. The unsupervised/supervised columns correspond to models with/without learning a sense-extensional semantic space. We observed that 1) all BERT-based models benefit from learning a sense-extensional semantic space, suggesting the presence of regularity shared among examples of sense extension across verb types; 2) both the prototype and exemplar WSG models consistently supersede other baseline models in both unsupervised and supervised setups, indicating that chaining mechanisms are useful inductive biases for modeling the generative processes of word meaning extension. [4]

Table 3 shows example predictions on 6 polysemous verbs made by a supervised prototype WSG model (the full model) and a supervised BERT-MLM baseline trained on the same datasets. The full model successfully predicts many types of sense extension, including weak metaphor (the *pass* example, where the location-type argument of a predicate is weakened to its super-type of scalar attribute), strong metaphor (the *throw* and *stretch* examples, where verbs with concrete noun arguments are extended to predicate abstract terms) and logical metonymy (the *appear* (in) example of event-for-place type coercion). In contrast, the MLM baseline exhibits a greater tendency to predict a "literal" paraphrase for a partitioned token (e.g., all of its top-3 predicted paraphrase tokens for "throw COGNITION" have senses from the same abstract domain). In addition, both chaining-based and baseline models still struggle in predicting some usages that involve strong non-literal sense extension (e.g., the *grasp* example).

6 Model interpretation and analysis

6.1 Sense relatedness and model predictability

Prior work in psycholinguistics suggests that both adults and children often find it easier 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). Here we investigate whether our WSG models exhibit human-like sensitivity to the conceptual relateness of source-target sense pairs in the full partition setup. We quantify the degree of conceptual relatedness for domain-based partitioned tokens by computing the difference in mean concreteness score (Brysbaert et al., 2014) of their attested object nouns, and for tokens representing synset-based senses, we take the Wu-Palmer semantic distance (Wu and Palmer, 1994) between their associated WordNet synsets as the similarity measurement. Figure 2 shows the performance of 4 WSG models by binning sense pairs based on their degrees of conceptual similarity. We observe that all models yield worse prediction for sense pairs that are conceptually more distant (e.g., metaphors), while generally performing better on pairs that are

[4] We also summarize results on a subset of the domain-based dataset where the target sense is more abstract than the source – see Appendix C.

Sense type	Model	Mean reciprocal rank (MRR-100)			
		Full partition		Leave-one-out partition	
		Unsupervised	Supervised	Unsupervised	Supervised
Domain-based	Random Baseline	5.2±0.0	–	5.2±0.0	–
	BERT-STS	13.7±1.2	33.1±2.4	38.5±1.9	62.5±1.6
	BERT-MLM	15.2±1.5	35.4±1.9	44.4±2.0	64.8±1.3
	WSG-Prototype	**19.8±1.3**	**52.3±0.3**	55.0±1.4	**86.9±1.6**
	WSG-Exemplar	19.5±1.1	47.9±1.5	**78.2±2.7**	86.6±1.1
Synset-based	Random Baseline	5.2±0.0	–	5.2±0.0	–
	BERT-STS	14.2±2.0	16.9±1.2	26.2±2.0	55.0±1.8
	BERT-MLM	17.3±0.8	19.1±0.7	25.7±0.8	59.9±0.5
	WSG-Prototype	**20.7±0.5**	**30.8±0.6**	36.0±1.1	**73.4±1.0**
	WSG-Exemplar	19.6±0.8	29.5±1.2	**68.0±2.9**	72.9±0.6

Table 2: Summary of model MRR-100 scores (%) for word sense generation in the two datasets. Numbers after ± are standard deviations over 5 sets of independently sampled negative candidate tokens.

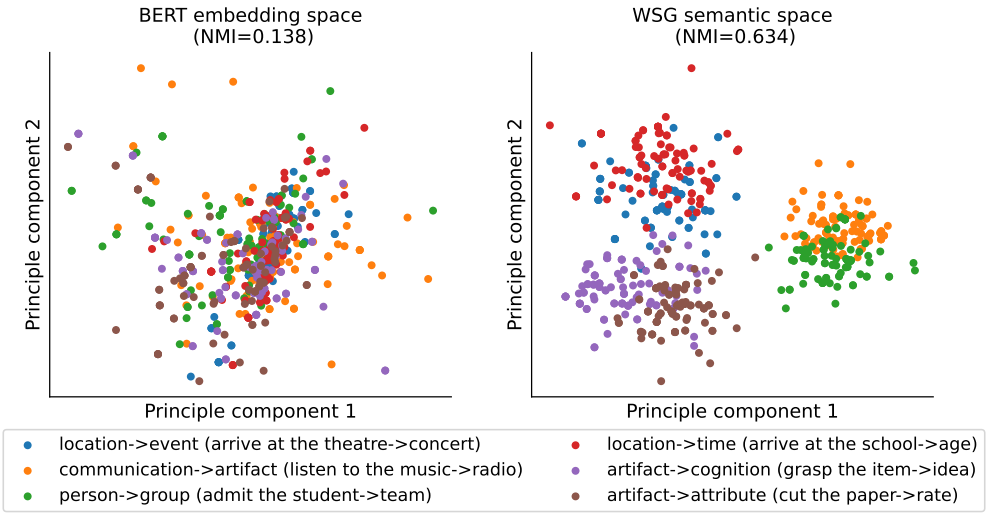


Figure 3: Principal component visualization of prototype difference embeddings of partitioned source-target token pairs, under BERT-MLM baseline (left) and prototype WSG model in sense-extensional semantic space (right).

conceptually more related (e.g., metonymy).

6.2 Interpreting the learned semantic space

To better understand the effect of the sense-extensional semantic space, we compute the mean Euclidean distance between the prototypes (i.e., global mean) of embedded usage sentences for all fully-partitioned source-target token pairs yielded by two types of chaining-based models. As shown in Table 4, both models benefit from learning the extensional semantic space by bringing closer novel and existing senses of the same word. Moreover, pushing partitioned tokens closer for verbs in the training set also results in a more compact embedded sense inventory for unseen verbs in the evaluation set, suggesting that the WSG models have captured some regularity shared across the meaning transformations of various word types.

To further interpret the information captured in the WSG models, we also performed unsupervised K-means clustering on prototype difference vectors of all fully partitioned domain-based source-target token pairs taken from 1) the embedding space of the BERT-MLM baseline and a sense-extensional semantic space learned by a prototype WSG model. We then compute the normalized mutual information (NMI) between the cluster labels yielded by the K-means algorithm and the ground-truth pairings of source-target domains for each example. Figure 3 shows the clustering results for two semantic spaces, together with their NMI scores against the ground-truth sense extension type labels. We observe that the learned space captures systematic sense extensional patterns by

Model	Top-3 source tokens predicted by model	Predicted rank
1.1 Verb: *throw*; target sense: **to throw a COGNITION**; true source sense: **to throw an ARTIFACT** Usage context: Constructors ... which returned null upon failure were changed to *throw* an exception instead.		
BERT-MLM	create COGNITION, develop COGNITION, provide COGNITION	55/100
WSG-Prototype	create COGNITION, throw ARTIFACT , build ARTIFACT	2/100
1.2 Verb: *pass*; target sense: **to pass an ATTRIBUTE**; true source sense: **to pass an ARTIFACT** Usage context: After 22 minutes of flight, the aircraft *passed* its assigned altitude.		
BERT-MLM	retain ATTRIBUTE, lose ATTRIBUTE, regain ATTRIBUTE	41/100
WSG-Prototype	pass ARTIFACT, enter LOCATION, return to ARTIFACT	1/100
1.3 Verb: *appear (in)*; target sense: **to appear in an EVENT**; true source sense: **to appear in a LOCATION** Usage context: He ... *appeared in* 24 league matches as well as United's FA Cup defeat to Burnley.		
BERT-MLM	host EVENT, achieve EVENT, lose EVENT	62/100
WSG-Prototype	come to GROUP, star in ACT, compete in ACT	6/100
2.1 Verb: *cover*; target sense: **be responsible for reporting (news)**; true source sense: **to form a cover over** Usage context: Generally, only reporters who *cover* breaking news are eligible.		
BERT-MLM	work (operate in a place), take (be a student), write (communicate by writing)	78/100
WSG-Prototype	practice (carry out as job), sponsor (be a client of), monitor (supervise someone)	10/100
2.2 Verb: *stretch*; target sense: **to extend the scope or meaning of**; true source sense: **to make longer by pulling** Usage context: ... the usage of a new "entwinement " standard ... *stretched* the doctrine beyond its permissible limits.		
BERT-MLM	spoil (alter), meet (satisfy), understand (comprehend)	59/100
WSG-Prototype	stem (remove the stem), stretch (make longer by pulling), extend (stretch to a greater length)	2/100
2.3 Verb: *grasp*; target sense: **to get the meaning of**; true source sense: **to hold firmly** Usage context: Madonna later acknowledged that she had not *grasped* the concept of her mother dying.		
BERT-MLM	appreciate (realize fully), understand (comprehend), enjoy (get pleasure from)	86/100
WSG-Prototype	understand (comprehend), resolve (understand the meaning of), read (interpret)	38/100

Table 3: Example novel sense predictions from the full prototype model and the BERT-MLM baseline (both trained on WSG) on domain-based dataset (first 3 examples) and synset-based dataset (last 3 examples).

Model	Mean Euclidean distance	
	Training	Testing
BERT	11.85±2.93	11.96±2.96
+ WSG-Prototype	1.14±0.38	1.67±0.43
+ WSG-Exemplar	1.21±0.34	1.82±0.50

Table 4: Mean Euclidean distance between the prototypes of source and target tokens.

Model	MLM perplexity	STS correlation
BERT	0.337	0.665
+WSG-Prototype	0.339	0.632
+WSG-Exemplar	0.352	0.634

Table 5: Performance on common NLP tasks for the vanilla BERT model and WSG models.

forming well-separated clusters for tokens pairs from the same extensional type.

6.3 Evaluation on common NLP tasks

We finally examined how learning WSG might affect a language model on common NLP tasks. Table 5 shows performance on two general NLP tasks for the BERT encoders with and without supervised learning on WSG: 1) masked langauge modeling (MLM) on Wikitext-103 dataset; and 2) semantic textual similarity (STS) prediction on the STS benchmark dataset (Cer et al., 2017). We found that contextualized language models trained on WSG largely preserved their linguistic competence on both tasks, suggesting that our WSG learning framework is not simply a fine-tuning technique designed for a specific task, but rather a more fundamental exercise of learning flexible embedding spaces that may improve the semantic generalizability of language models.

7 Conclusion

We have presented a framework of word sense generation that supports lexical items to spawn new senses in scenarios that involve novel context. Our results show that chaining provides a general mechanism for extending to novel senses including metaphor and metonymy, and learning a transformed sense-extensional space enables systematic generalization in word sense extension. In contrast with the established traditions in word sense disambiguation and induction, our work emphasizes a generative approach to model word senses that is scalable and applicable to natural sentences and a broad set of words in the lexicon. Future work may extend our framework to incorporate temporal dimensions, other word classes and languages.

8 Ethical considerations

In this section, we discuss the limitations and potential risks of our work.

8.1 Limitations

Our work has some limitations. First, the current study considers sense generation for English verbs, and it therefore does not account for all common types of sense extension in English (e.g., English noun-to-verb conversion, as discussed in Clark and Clark 1979; Yu and Xu 2022) and other languages. Future work can consider extending our WSG framework more broadly to other word classes and languages.

Second, our framework does not explicitly consider the temporal order via which word senses have emerged. In particular, in the data collection step, we choose source-target sense pairs based on random partitioning of word sense inventories, whereas an alternative 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 possible source-target sense pairs against sampled alternatives, which is a potential area for future extension.

8.2 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 WSG 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 WSG models cross-linguistically to cover various innovative lexical uses should help alleviate this issue.

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.

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

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

Tuhin Chakrabarty, Xurui Zhang, Smaranda Muresan, and Nanyun Peng. 2021. Mermaid: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4250–4261.

Eve V Clark and Herbert H Clark. 1979. When nouns surface as verbs. *Language*, pages 767–811.

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.

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

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).

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.

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

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

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. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

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.

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.

Jey Han Lau, Paul Cook, Diana McCarthy, Spandana Gella, and Timothy Baldwin. 2014. Learning word sense distributions, detecting unattested senses and identifying novel senses using topic models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 259–270.

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

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.

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.

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.

Gregory L Murphy. 1997. Reasons to doubt the present evidence for metaphoric representation. *Cognition*, 62(1):99–108.

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.

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.

Giulia Rambelli, Emmanuele Chersoni, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2020. Comparing probabilistic, distributional and transformer-based models on logical metonymy interpretation. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*.

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.

Mahesh Srinivasan and Susan Carey. 2010. The long and the short of it: On the nature and origin of functional overlap between representations of space and time. *Cognition*, 116(2):217–241.

Kevin Stowe, Tuhin Chakrabarty, Nanyun Peng, Smaranda Muresan, and Iryna Gurevych. 2021. Metaphor generation with conceptual mappings. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6724–6736.

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

Paul Thibodeau and Frank H Durgin. 2008. Productive figurative communication: Conventional metaphors facilitate the comprehension of related novel metaphors. *Journal of Memory and Language*, 58(2):521–540.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686.

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.

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.

Lei Yu and Yang Xu. 2022. Noun2Verb: Probabilistic Frame Semantics for Word Class Conversion. *Computational Linguistics*, pages 1–36.

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.

A Implementations of WSG models

We use the bert-base-uncased configuration provided by Hugging Face (Wolf et al., 2020) to initialize all BERT-based WSG models (two baselines and two chaining-based models). During MLM pretraining of BERT models to learn novel partitioned tokens, we randomly mask 15% of tokens in each sentence, and train each model on predicting the masked tokens. 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, all chaining-based models are trained on the objective function in Eq.3.3 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). Two baseline models are trained on the same set of usage data using Eq.2 and Eq.3 respectively, with an Adam learning rate of 2e-5 and a batch size of 32, for 4 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 25 minutes per epoch for MLM pretraining, and 10 minutes per epoch for sense-extensional semantic space learning.

B Description of supersense-based domains

WordNet organize noun synsets into 26 supersense domains based on their semantic coherence. During data collection, we first run the WSD model of (Bevilacqua and Navigli, 2020) on usage sentences to identify the WordNet synset label for each noun object, and then tag it with the associated supersense label. We then take the top-11 most frequent supersenses with at least 10,000 identified noun objects to construct our domain-based WSG dataset. Table 6 shows detailed information about the 11 selected supersense categories.

C Results on WSG examples with decreasing sense concreteness

Research in cognitive science suggests that concrete and embodied word senses tend to be extended to more abstract senses to achieve better communication efficiency and learnability (Murphy, 1997; Srinivasan and Carey, 2010; Thibodeau and Durgin, 2008). We therefore also compute performance scores of our WSG models on subsets of the domain-based usage dataset where the meaning of a verb is extended from concrete senses to a more abstract one (e.g. the metaphorical extensions of *arrive* as shown in Figure 1). In particular, we define a source-target sense pair to be an example of concrete-to-abstract sense extension if the average object noun concreteness score of the target token is lower than that of its source sibling. Table C summarizes the results, from which we observe that almost all models perform slightly worse on concrete-to-abstract extensions compared to the general setup, while still making significantly better predictions than the random baseline.

Supersense	Definition	Sample nouns
noun.act	nouns denoting acts or actions	attempt, performance, exchange
noun.artifact	nouns denoting man-made objects	aircraft, phone, guitar
noun.attribute	nouns denoting attributes of people and objects	popularity, style, power
noun.cognition	nouns denoting cognitive processes and contents	viewpoint, imagination, scheme
noun.communication	nouns denoting communicative processes and contents	proposal, screenplay, film
noun.event	nouns denoting natural events	tournament, final, competition
noun.group	nouns denoting groupings of people or objects	family, league, party
noun.location	nouns denoting spatial position	province, area, neighborhood
noun.person	nouns denoting people	officer, visitor, mother
noun.state	nouns denoting stable states of affairs	existence, friendship, injury
noun.time	nouns denoting time and temporal relations	era, day, season

Table 6: Definitions and sample members of the WordNet noun supersenses used to create the domain-based WSG dataset.

Model	Mean reciprocal rank (MRR-100)			
	Full partition		Leave-one-out partition	
	Unsupervised	Supervised	Unsupervised	Supervised
BERT-STS	11.5±1.2	34.6±2.0	37.7±1.8	61.6±1.6
BERT-MLM	13.6±1.4	35.8±1.7	43.0±2.2	64.1±0.9
BERT-Prototype	17.5±1.1	49.9±0.5	51.6±1.5	81.4±1.9
BERT-Exemplar	17.8±1.1	45.8±1.4	75.3±2.9	79.7±1.0

Table 7: Summary of model MRR-100 scores (%) on the subset of domain-based word sense generation dataset where the source sense has higher concreteness score than its target sibling token (i.e. extensions from concrete to abstract senses). Numbers after ± are standard deviations over 5 sets of independently sampled negative candidate tokens.