Predicate Representations and Polysemy in VerbNet Semantic Parsing

James Gung*
University of Colorado Boulder
james.gung@colorado.edu

Martha Palmer
University of Colorado Boulder
martha.palmer@colorado.edu

Abstract

Despite recent advances in semantic role labeling propelled by pre-trained text encoders like BERT, performance lags behind when applied to predicates observed infrequently during training or to sentences in new domains. In this work, we investigate how semantic role labeling performance on low-frequency predicates and out-of-domain data can be improved by using VerbNet, a verb lexicon that groups verbs into hierarchical classes based on shared syntactic and semantic behavior and defines semantic representations describing relations between arguments. We find that VerbNet classes provide an effective level of abstraction, improving generalization on low-frequency predicates by allowing them to learn from the training examples of other predicates belonging to the same class. We also find that joint training of VerbNet role labeling and predicate disambiguation of VerbNet classes for polysemous verbs leads to improvements in both tasks, naturally supporting the extraction of VerbNet's semantic representations.

1 Introduction

Semantic role labeling (SRL) is a form of shallow semantic parsing that involves the extraction of predicate arguments and their assignment to consistent roles with respect to the predicate, facilitating the labeling of e.g. *who* did *what* to *whom* (Gildea and Jurafsky, 2000). SRL systems have been broadly applied to applications such as question answering (Berant et al., 2014; Wang et al., 2015), machine translation (Liu and Gildea, 2010; Bazrafshan and Gildea, 2013), dialog systems (Tur and Hakkani-Tür, 2005; Chen et al., 2013), metaphor detection (Stowe et al., 2019), and clinical information extraction (Gung, 2013; MacAvaney et al., 2017). Recent approaches to SRL have achieved large gains in performance through the use of pre-trained text encoders like ELMo and BERT (Peters et al., 2018; Devlin et al., 2019). Despite these advances, performance on low-frequency predicates and out-of-domain data remains low relative to in-domain performance on higher frequency predicates.

	Billy	***consoled***	*the puppy*
PB	Arg0	console.01	Arg1
VN	Stimulus	amuse-31.1	Experiencer
	Billy	***walked***	*the puppy*
PB	Arg0	walk.01	Arg1
VN	Agent	run-51.3.2-2-1	Theme

Table 1: Comparison of PropBank (PB) and VerbNet (VN) roles for predicates *console* and *walk*. VerbNet's thematic role assignments (e.g. Stimulus vs. Agent and Experiencer vs. Theme) are more dependent on the predicate than PropBank's numbered arguments.

The assignment of role labels to a predicate's arguments is dependent upon the predicate's sense. PropBank (Palmer et al., 2005) divides each predicate into one or more *rolesets*, which are coarse-grained sense distinctions that each provide a set of core numbered arguments (A0-A5) and their corresponding definitions. VerbNet (VN) groups verbs into hierarchical *classes*, each class defining a set of valid syntactic frames that define a direct correspondence between thematic roles and syntactic realizations, e.g. *Agent REL Patient* (e.g. *John broke the vase*) or *Patient REL* (e.g. *The vase broke*) for *break-45.1* (Schuler, 2005).

Recent PropBank (PB) semantic role labeling models have largely eschewed explicit predicate disambiguation in favor of direct prediction of semantic roles in end-to-end trainable models (Zhou and Xu, 2015; He et al., 2017; Shi and Lin, 2019).

*Work done prior to joining Amazon.

Proceedings of the 14th International Conference on Computational Semantics, pages 51–62
June 17–18, 2021.

This is possible for several reasons: First, PropBank's core roles and modifiers are shared across all predicates, allowing a single classifier to be trained over tokens or spans. Second, although definitions of PB roles are specific to the different senses of each predicate, efforts are made when creating rolesets to ensure that A0 and A1 exhibit properties of Dowty's prototypical Agent and prototypical Patient respectively (1991). Finally, PB rolesets are defined based on VN class membership, with predicates in the same classes thus being assigned relatively consistent role definitions (Bonial et al., 2010).

Unlike PropBank, VerbNet's thematic roles are shared across predicates and classes with consistent definitions. However, VN roles are more dependent on the identity of the predicate (Zapirain et al., 2008; Merlo and Van Der Plas, 2009). Examples of PropBank and VerbNet roles illustrating this are given in Table 1. Consequently, VN role labeling models may benefit more from predicate features than PropBank. Furthermore, while it is possible to identify PB or VN roles without classifying predicate senses, linking the resulting roles to their definitions or to the syntactic frames and associated semantic primitives in VN does require explicit predicate disambiguation (Brown et al., 2019). Therefore, predicate disambiguation is often an essential step when applying SRL systems to real-world problems.

In this work, we evaluate alternative approaches for incorporating VerbNet classes in English VerbNet and PropBank role labeling. We propose a joint model for SRL and VN predicate disambiguation (VN classification), finding that joint training leads to improvements in VN classification and role labeling for out-of-domain predicates. We also evaluate VN classes as predicate-specific features. Using gold classes, we observe significant improvements in both PB and VN SRL. We also observe improvements in VN role labeling when using predicted classes and features that incorporate all valid classes for each predicate[1].

2 Background and Related Work

VerbNet VerbNet is a broad-coverage lexicon that groups verbs into hierarchical classes based on shared syntactic and semantic behavior (Schuler, 2005). Each VN class is assigned a set of thematic roles that, unlike PB numbered arguments, maintain consistent meanings across different verbs and classes. VN classes provide an enumeration of syntactic frames applicable to each member verb, describing how the thematic roles of a VN class may be realized in a sentence. Every syntactic frame entails a set of low-level semantic representations (primitives) that describe relations between thematic role arguments as well as changes throughout the course of the event (Brown et al., 2018). The close relationship between syntactic realizations and semantic representations facilitates straightforward extraction of VN semantic predicates given identification of a VN class and corresponding thematic roles. VN primitives have been applied to problems such as machine comprehension (Clark et al., 2018) and question generation (Dhole and Manning, 2020).

Comparing VerbNet with PropBank Yi et al. (2007) use VN role groupings to improve label consistency across verbs by reducing the overloading of PropBank's numbered arguments like A2. Comparing SRL models trained on PB and VN, Zapirain et al. (2008) find that their VerbNet model performs worse on infrequent predicates than their PB model, and suggest that VN is more reliant on the identity of the predicate than PB based on experiments removing predicate-specific features from their models. They suggest that the high consistency of A0 and A1 enables PB to generalize better without relying on predicate-specific information.

Merlo and Van Der Plas (2009) provide an information-theoretic perspective on the comparison of PropBank and VerbNet, demonstrating how the identity of the predicate is more important to VN SRL than for PB by comparing the conditional entropy of roles given verbs as well as the mutual information of roles and verbs. In multilingual BERT probing studies comparing several SRL formalisms, Kuznetsov and Gurevych (2020) find that layer utilization for predicates differs between PB and VN. PB emphasizes the same layers used for syntactic tasks, while VN uses layers associated with tasks used more prevalently in lexical tasks. These findings reinforce the importance of predicate representations to VerbNet.

SRL and Predicate Disambiguation Previous work has investigated the interplay between predicate sense disambiguation and SRL. Dang and Palmer (2005) improve verb sense disambiguation

[1]Our code is available at https://github.com/jgung/verbnet-parsing-iwcs-2021.

(VSD) using features based on semantic role labels. Moreda and Palomar (2006) find that explicit verb senses improve PB SRL for verb-specific roles like A2 and A3, but hurt on adjuncts. Yi (2007) find that using gold standard PB roleset IDs as features in an SRL model improves performance only on highly polysemous verbs. Dahlmeier et al. (2009) propose a joint probabilistic model for *preposition* disambiguation and SRL, finding an improvement over independent models.

Predicate disambiguation plays a critical role in FrameNet (Baker et al., 1998) parsing, in part because FrameNet's role inventory is more than an order of magnitude larger than that of PB and VN. This richer, more granular role inventory lends advantages to approaches that constrain role identification to the set of valid roles for the predicted frame (Das et al., 2014; Hermann et al., 2014), or that jointly encode argument and role representations given identified frames (FitzGerald et al., 2015).

LM Pre-training and SRL Language model (LM) pre-training has become ubiquitous in natural language processing tasks, with LM encoders like ELMo propelling forward the state of the art in SRL (Peters et al., 2018). We are interested in whether a strong baseline model using a LM encoder such as BERT can be further improved by incorporating external knowledge from lexical resources like VN.

BERT (Devlin et al., 2019) is a Transformer encoder (Vaswani et al., 2017) jointly trained using two objectives: a masked language modeling objective to predict the identity of randomly-masked tokens in the input, as well as a next sentence prediction task (NSP) intended to encourage the model to encode the relationship between sentence pairs (henceforth referred to as *Sent. A* and *Sent. B*). Sentences are tokenized using WordPiece (Wu et al., 2016). As a Transformer encoder, BERT applies multiple layers of a multi-headed self-attention mechanism to progressively build contextual token-level representations. In our experiments, we use encodings from the final layer.

3 Semantic Role Labeling with BERT

Our baseline SRL model closely follows Shi and Lin (2019). We thus approach SRL as a sequence tagging task, predicting per-word, IOB-encoded (**I**n, **O**ut, **B**egin) role labels independently for each predicate in a sentence. A predicate-aware encoding of a sentence is produced using the target predicate as the *Sent. B* input to BERT. For example, the sentence *I tried opening it* is processed as:

CLS I tried opening it SEP opening SEP

for the verb *open*. This enables BERT to incorporate the identity of the predicate in the encoding of each word while clearly delineating it from tokens in the original sentence.

To simplify notation, we'll treat $\mathbf{LM}(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^{T_a \times D_{\text{LM}}}$ as shorthand for the final layer BERT encoding for a pair of sentences $\mathbf{a} = w_1, ..., w_{T_a}$ and $\mathbf{b} = w_1, ..., w_{T_b}$ with T_a and T_b words respectively, where D_{LM} gives BERT's hidden size. This is produced by applying WordPiece tokenization (**WP**) to each word in each sentence and concatenating the resulting sequences of token IDs with standard BERT-specific IDs:

$$\mathbf{w} = \left[\text{CLS}, \mathbf{WP}(\mathbf{a}), \text{SEP}, \mathbf{WP}(\mathbf{b}), \text{SEP}\right]$$

The resulting sequence of tokens $\mathbf{w}$ is encoded using BERT. We use the final layer outputs, taking vectors only for the first WordPiece token for each original word in *Sent. A* ($\mathbf{a}$), filtering out vectors corresponding to *Sent. B* ($\mathbf{b}$), SEP or CLS. The resulting matrix consists of a vector per word in *Sent. A*, avoiding any discrepancies between IOB-encoded word-level output labels and WordPiece tokens used as inputs.

Following previous work (Zhou and Xu, 2015; He et al., 2017), we use a marker feature as an indicator for the specific location of the predicate within the sentence. For a sentence, $w_1, ..., w_T$, with a predicate given by index $p \in 1...T$, we compute a predicate-aware, contextualized embedding $\mathbf{x}_{pt}$ of each word as

$$\mathbf{x}_{pt} = \left[\mathbf{LM}(w_{1...T}, w_p)_{(t)}; \mathbf{W}^{(mark)}_{(t=p)}\right] \quad (1)$$

with $\mathbf{W}^{(mark)} \in \mathbb{R}^{2 \times D_{mark}}$ and $\mathbf{x}_{pt} \in \mathbb{R}^{D_{\text{LM}}+D_{mark}}$, where D_{mark} provides the size of the predicate marker embedding.

The predicate's positional information from the marker is integrated using a bidirectional LSTM (Hochreiter and Schmidhuber, 1997), concatenating the hidden states for the forward and backward LSTMs at each timestep (omitting the p from $\mathbf{x}_{pt}$ for brevity):

$$\begin{aligned} \mathbf{h}^{(fw)}_t &= \text{LSTM}^{(fw)}(\mathbf{x}_{1...T})_{(t)} \\ \mathbf{h}^{(bw)}_t &= \text{LSTM}^{(bw)}(\mathbf{x}_{T...1})_{(T-t)} \\ \mathbf{h}^{(fb)}_t &= \left[\mathbf{h}^{(fw)}_t; \mathbf{h}^{(bw)}_t\right] \end{aligned} \quad (2)$$

The BiLSTM output at each timestep t is concatenated with that of the predicate's timestep and passed through a sequentially-applied linear transformation followed by a leaky ReLu ($\alpha = 0.1$):

$$\mathbf{x}_{pt}^{(mlp)} = \sigma\left(\mathbf{W}^{(mlp)}\left[\mathbf{h}_t^{(fb)}; \mathbf{h}_p^{(fb)}\right] + \mathbf{b}^{(mlp)}\right) \tag{3}$$

We apply a final linear projection from $\mathbf{x}_{pt}^{(mlp)}$ to IOB-encoded role labels:

$$\mathbf{s}_{pt}^{(srl)} = \mathbf{W}^{(srl)}\mathbf{x}_{pt}^{(mlp)} + \mathbf{b}^{(srl)} \tag{4}$$

where $\mathbf{s}_{pt}^{(srl)} \in \mathbb{R}^K$ provides the unnormalized scores for each of K possible role labels, with the probability of predicting a label for a given token t and predicate p given by:

$$P(y_{pt}^{(srl)}|w_{1...T}, w_p) = \text{softmax}(\mathbf{s}_{pt}^{(srl)}) \tag{5}$$

Like He et al. (2017), we apply constrained Viterbi decoding to restrict inferred label sequences to produce valid IOB sequences.

4 VerbNet Classes as Predicate Features

Verbs belonging to the same VN class share syntactic and semantic properties and the same set of thematic roles and syntactic frames. Replacing a predicate in a sentence with a different verb from the same class typically produces a syntactically coherent sentence and does not impact the proposition's thematic role labels. VN classes may thus provide an effective level of abstraction for predicates in SRL.

We hypothesize that using VN classes as predicate-specific features may help reduce sparsity issues for low-frequency and out-of-vocabulary (OOV) verbs. Intuitively, training examples for each member verb within a class contribute to the estimation of parameters associated with all other members of the same class, enabling the fine-tuning of predicate-level features even for OOV predicates. For example, a verb like *traipse* may rarely or never occur during training, but may belong to a class which appears hundreds of times in the form of more common verbs like *run* or *rush*. We investigate whether by sharing parameter updates across VN members, we can further improve generalization on infrequent verbs.

Methodology Intuitively, BERT's NSP pre-training task encourages some level of focus on *Sent. B* tokens from attention heads when processing tokens in *Sent. A*. The predicate feature presented by Shi and Lin (2019) and applied in our baseline model uses the predicate token as the *Sent. B* input to BERT and thus allows the encodings of tokens in a sentence to be conditioned directly on the predicate.

We propose to include tokens corresponding to the predicate's VN class as additional features as part of *Sent. B*. To realize this, we concatenate the corresponding VN class ID to *Sent. B* along with the predicate, updating the inputs given in Equation 1:

$$\mathbf{LM}(w_{1...T}, w_p w_s) \tag{6}$$

where w_s is a token corresponding to the VN class of the predicate w_p[2].

VerbNet Classification VN classes can be predicted automatically using a word sense disambiguation system. We propose a simple model for VerbNet classification: fine tune a pre-trained BERT encoder by applying a feedforward multi-layer perceptron (MLP) classifier over all VN classes to the BERT encoding associated with the first WordPiece of the target predicate.

We again condition BERT on the target predicate by including it as a feature (w_p) in *Sent. B*:

$$\begin{aligned} \mathbf{x}_p &= \mathbf{LM}(w_{1...T}, w_p)_{(p)} \\ \mathbf{x}_p^{(mlp)} &= \sigma\left(\mathbf{W}^{(mlp)}\mathbf{x}_p + \mathbf{b}^{(mlp)}\right) \\ \mathbf{s}_p^{(vncls)} &= \mathbf{W}^{(vncls)}\mathbf{x}_p^{(mlp)} + \mathbf{b}^{(vncls)} \end{aligned} \tag{7}$$

where $\mathbf{W}^{(vncls)} \in \mathbb{R}^{D_{mlp} \times V}$ projects over all V VN classes for all predicates. The probability for predicting a VN class $y_p^{(vncls)}$ for a given predicate and sentence is given by:

$$P(y_p^{(vncls)}|w_{1...T}, w_p) = \text{softmax}(\mathbf{s}_p^{(vncls)}) \tag{8}$$

This *single classifier* formulation is possible for lexicons like VN and FrameNet in which predicates share senses from a global sense inventory. While individual predicates have a specific set of valid senses, their senses are shared from the global lexicon. Kawahara and Palmer (2014) demonstrate

[2]In preliminary experiments, we found that directly modifying *Sent. A* drastically reduces the performance of the model and slows convergence.

that a single classifier approach to VN classification achieves competitive performance when using shared semantic features. Intuitively, by training the classifier across multiple verbs, the model parameters specific to each sense receive more updates, with infrequent verb-class pairs also benefiting from the examples of other verbs within the same class. At inference time, we constrain sense predictions to predicate-sense combinations observed in the training data, selecting the highest-scoring valid sense given the predicate. We evaluate models using both predicted and gold (ground truth) classes for w_s as PREDICTED CLASS and GOLD CLASS respectively.

VerbNet Classes without Disambiguation Like SRL, VerbNet classification accuracy declines in the long tail of low frequency senses and predicates. For this reason, incorrect sense predictions may negate the benefits of VN class features on precisely the instances for which they might be expected to be beneficial: OOV or rare predicates.

To avoid this problem while still retaining the benefits of parameter sharing for low frequency predicates with higher-frequency predicates belonging to the same VN class, we propose including the set of all possible classes for a given predicate as *Sent. B* features. To incorporate multiple senses, we simply concatenate them sequentially to *Sent. B*:

$$\mathbf{LM}(w_{1...T}, w_p w_{s_{1...k}}) \tag{9}$$

This allows the BERT encoder to attend over all possible VerbNet classes for a given predicate and sentence, without making a discrete decision about which class is correct. The extent and way in which the model incorporates the *Sent. B* tokens associated with the available classes is learned during training. The inputs to this model, later referred to as ALL CLASSES are identical to PREDICTED CLASS and GOLD CLASS models for monosemous predicates.

5 Joint VerbNet Classification and SRL

Features that are useful for SRL may also be useful in predicting the sense of a predicate. For example, surface-level syntactic awareness that the argument of a predicate is a clause instead of a noun phrase may change the expected sense of a verb (bring-11.3 vs. characterize-29.2):

Bob *took* Mary to the doctor.
John *took* Mary to be a doctor.

The semantic classes of arguments are also often important in determining the sense of a given predicate (dub-29.3.2 vs. get-13.5.1):

John *called* Mary a name.
John *called* Mary a car.

This dependency between SRL and predicate sense disambiguation together with the prevalence of shared features between the two tasks makes them a good candidate for multi-task learning (Caruana, 1998).

Multi-task Model Much of recent work in multi-task learning for SRL has focused on syntactic tasks such as syntactic parsing as auxiliary objectives (Strubell et al., 2018; Swayamdipta et al., 2018; Xia et al., 2019; Zhou et al., 2020). We first investigate an MTL approach that predicts semantic role labels and predicate senses independently given a shared BERT encoder. We extend our baseline SRL model, adding an additional *head* that is trained to predict the target predicate's sense, as described in Equation 8. The negative log likelihood of a single training instance with predicate p and token sequence $\mathbf{x} = w_{1...T}$ with T tokens is then given by:

$$-\sum_{t=1}^{T}\Big[\log P(y_{pt}^{(srl)}|\mathbf{x}, p)\Big] + \lambda_{vncls}\log P(y_p^{(vncls)}|\mathbf{x}, p) \tag{10}$$

with λ_{vncls} weighting the contribution of VerbNet class prediction to the overall objective. For brevity, we henceforth refer to this model as SRL + VSD.

We also investigate conditioning role labeling directly on predicted predicate senses. We implement this by concatenating a weighted label embedding of the target predicate's predicted class to each of the SRL head's input vectors, $\mathbf{x}_{pt}^{(srl)}$. To compute the weighted label embedding of a given VN class $y_p^{(vncls)}$ we follow Hashimoto et al. (2017):

$$\mathbf{y}_p^{(vncls)} = \sum_{k=1}^{K} P(y_p^{(vncls)} = k|\mathbf{x}, p)\mathbf{W}_{(k)}^{(vncls)} \tag{11}$$

with $\mathbf{W}^{(vncls)} \in \mathbb{R}^{K \times D_{vncls}}$ and $\mathbf{y}_p^{(srl)} \in \mathbb{R}^{D_{vncls}}$. The input to the SRL head is then given by:

$$\mathbf{x}_{pt}^{(srl)} = \big[\mathbf{LM}(w_{1...T}, w_p)_{(t)}; \mathbf{W}_{(t=p)}^{(mark)}; \mathbf{y}_p^{(vncls)}\big] \tag{12}$$

VerbNet class embeddings are initialized using the average of word embeddings corresponding to members of each class. During training, we use embeddings of predicted labels to avoid a discrepancy between the inputs to the SRL head between training and inference, when the gold labels are no longer available. In preliminary experiments, we used gold labels, similar to *teacher forcing* as described in Williams and Zipser (1989), but found that performance degraded when applied to predicted labels. We refer to the model described in this section as SRL | VSD.

6 Experiments

All models are implemented using Tensorflow 1.13 (Abadi et al., 2016) and are trained on a single NVIDIA GTX 1080 Ti GPU. We use the 110M parameter cased BERT-Base model available in Tensorflow Hub[3], with $D_{\text{LM}} = 768$. To align with Shi and Lin (2019), D_{mark} is set to 10, and LSTM and MLP hidden state sizes are set to 768 and 300 respectively. Dropout rates of 0.1 are applied to BERT outputs as well as after ReLu transforms in MLPs. Recurrent dropout (Gal and Ghahramani, 2016) with a rate of 0.1 is applied in LSTMs on hidden states and outputs. To initialize VerbNet class embeddings, we use 100-dimensional GloVe embeddings (Pennington et al., 2014) averaged over member verbs ($D_{vncls} = 100$). λ_{vncls} is set to 0.5 after a preliminary search over $\{0.1, 0.5, 1.0\}$.

We follow the fine-tuning methodology described in Devlin et al. (2019), using Adam (Kingma and Ba, 2014) with a batch size of 16. The learning rate is warmed up linearly from 0 to 5e-5 for 10% of training, then decayed linearly to 0 for the rest of training. Models are trained for up to 8 epochs. The best-performing checkpoint on the development set, evaluated at every half epoch, is selected for evaluation.

Unless otherwise mentioned, we train and evaluate all models with at least 7 independent random initializations, and present mean scores in our comparisons. To establish statistical significance, we apply a test for *Almost Stochastic Dominance* (Dror et al., 2019) between test score distributions, using $\alpha = 0.05$. Numbers in bold indicate highest average performance within a given evaluative setting, with a single star indicating statistical significance of almost stochastic dominance over our baseline models for each experiment, and two stars indicating stochastic dominance ($\epsilon = 0$). For example, a value in a table of $\mathbf{88.2}^{\star\star}{}_{\pm 0.2}$ indicates that a model has a mean test score (e.g. F_1 or accuracy) of 88.2, with a standard deviation of 0.2, and is stochastically dominant over the baseline.

[3] https://tfhub.dev/google/bert_cased_L-12_H-768_A-12/1

System	CoNLL-2005 WSJ	CoNLL-2005 Brown	CoNLL-2012 Test
Peters et al. (2018)	-	-	84.6
He et al. (2018)	87.4	80.4	85.5
Ouchi et al. (2018)	87.6	78.7	86.2
Li et al. (2019)	87.7	80.5	86.0
Shi and Lin (2019)	88.1	80.9	86.2
Our Baseline	$87.5_{\pm 0.2}$	$81.2_{\pm 0.4}$	$86.2_{\pm 0.1}$

Table 2: Comparison of baseline SRL system on CoNLL-2005 and CoNLL-2012 against models applying pre-trained encoders of comparable size (F_1).

Datasets We use English PropBank datasets from CoNLL-2005 (Carreras and Màrquez, 2005) and the CoNLL-2012 split (Pradhan et al., 2013) for OntoNotes (Hovy et al., 2006) in order to situate our baseline mode among recent work in PB SRL. We compare against models of similar size (120M parameters) with pre-identified predicates.

The SemLink corpus (Palmer, 2009) is currently the only dataset that contains explicit VerbNet thematic role annotations with VN sense annotations. SemLink contains mappings between VN, PB and FrameNet, with annotations performed over a subset of the CoNLL-2005 PB WSJ annotations and Brown corpus out-of-domain test set (Carreras and Màrquez, 2005). Using SemLink thus allows us to evaluate performance for both PB and VN roles on the same source text. Following Zapirain et al. (2008), we restrict evaluation to propositions with PB core arguments fully mapped to VN thematic roles. This accounts for 56% of the original corpus. We include PB modifier roles in addition to VN thematic roles.

Baseline Comparisons Our baseline SRL model achieves comparable performance to Shi and Lin (2019) on both CoNLL-2012 and CoNLL-2005 and thus has performance similar to state-of-the-art models of the same size.

To compare our VerbNet classification models against prior work, we train and evaluate a publicly available state-of-the-art VN classification system directly on the SemLink corpus. We use Clear-

WSD[4], which is a sense disambiguation library tailored for verb sense disambiguation based on linear models over features constructed from an ensemble of word representations applied over syntactic relations (Palmer et al., 2017).

VerbNet Models The results of our experiments are shown in Table 3. First, we find that incorporating gold VerbNet classes (GOLD CLASS) significantly improves VerbNet SRL, providing a 15% relative error reduction on out-of-domain data (80.1 to 83.0), and 6% reduction on in-domain data (87.4 to 88.2). In PB SRL, gold classes are also beneficial, but to a lesser degree. ALL CLASSES and PREDICTED CLASS models improve both in-domain and out-of-domain VN SRL.

Predicting both VN classes and semantic roles from a single encoder reduces the total computational resources required to make predictions from separate models, providing a practical benefit. Additionally, we are interested in determining whether our multi-task models lead to improvements in generalization. Our multi-task model SRL + VSD, which does not condition thematic role prediction on predicted senses, does not have a significant effect on VN SRL performance. However, we do find that conditioning SRL on VN class predictions in a multi-task model (SRL | VSD) leads to a significant improvement in performance on the out-of-domain Brown test set for VN SRL. No significant change is observed on the in-domain WSJ test set, or when the model is applied to PB SRL.

We also evaluate the impact of multi-task learning on predicate disambiguation (VN classification). First, we find that even our baseline model is competitive with the highly-specialized approach for verb sense disambiguation provided in ClearWSD (Table 4). Comparing our joint VN SRL models with a single task baseline for VN classification, we observe a significant improvement on WSJ test data when incorporating multi-task supervision from SRL. This approach is related to earlier use of SRL features for verb sense disambiguation reported in Dang and Palmer (2005), and the positive result is consistent with their findings.

7 Analysis

Monosemous vs. Polysemous Predicates To understand the impact of VerbNet class features, we break down our evaluation by polysemous and monosemous verbs in Table 5. First, we observe that incorporating VN classes improves F_1 scores for monosemous verbs in both models. This is expected, as monosemous verbs are typically lower frequency, with low-frequency and OOV verbs benefiting the most from parameter sharing with other verbs belonging to the same VN classes. We also observe a significant improvement on polysemous verbs in the WSJ (in-domain) test set when including VN features. However, polysemous verbs in the Brown (out-of-domain) test set only benefit from using explicitly predicted classes, but not when using all valid classes for each predicate.

Why does ALL CLASSES improve performance on out-of-domain data for monosemous verbs, but not polysemous verbs? Intuitively, the per-verb distributions of VerbNet classes may change considerably between two domains. Using a correctly-predicted class may help mitigate errors on verbs for which one class was dominant during training, but a different class or set of classes are observed during testing in the new domain. This benefit would not be observed with ALL CLASSES as for a given verb, the same classes used as model inputs during training would be used as inputs on out-of-domain data. However, VN classes receive fewer updates during training when using only predicted classes. Thus, verbs appearing in classes that never or rarely appeared during training will not benefit from PREDICTED CLASS features. ALL CLASSES may mitigate this issue, since even if a specific class does not appear in the training data, it still can receive updates from examples of polysemous member verbs that belong to other classes (and improved performance over PREDICTED CLASS on monosemous verbs on the out-of-domain Brown test set supports this). As future work, a promising direction may therefore be to combine PREDICTED CLASS and ALL CLASSES features.

Out-of-Vocabulary Predicates How well do models incorporating VerbNet features generalize on out-of-vocabulary and rare predicates? We split an evaluation on the WSJ development set into 5 bins by training set predicate frequency (shown in Figure 1). Comparing development F_1 scores for ALL CLASSES and PREDICTED CLASS models against our baseline model, we note that VN classes improve SRL performance most for predicates appearing 0-50 times in the training data, which account for 24.4% of instances in the development set.

[4] https://github.com/clearwsd/clearwsd

System	PropBank WSJ	PropBank Brown	VerbNet WSJ	VerbNet Brown
Zapirain et al. (2008)	$78.9_{\pm 0.9}$	–	$77.0_{\pm 0.9}$	$62.9_{\pm 1.0}$
Baseline	$88.5_{\pm 0.1}$	$82.4_{\pm 0.5}$	$87.4_{\pm 0.2}$	$80.1_{\pm 0.4}$
SRL + VSD	$88.2_{\pm 0.2}$	$82.8^{\star}_{\pm 0.6}$	$87.3_{\pm 0.1}$	$80.0_{\pm 0.7}$
SRL | VSD	$88.3_{\pm 0.2}$	$82.2_{\pm 0.4}$	$87.4_{\pm 0.2}$	$80.6^{\star\star}_{\pm 0.4}$
PREDICTED CLASS	$88.3_{\pm 0.1}$	$81.2_{\pm 0.6}$	$87.6^{\star\star}_{\pm 0.1}$	$80.9^{\star\star}_{\pm 0.6}$
ALL CLASSES	$\mathbf{88.6}^{\star}_{\pm 0.3}$	$82.3_{\pm 0.5}$	$\mathbf{87.6}^{\star\star}_{\pm 0.2}$	$\mathbf{81.1}^{\star\star}_{\pm 0.6}$
GOLD CLASS	$\mathbf{88.7}^{\star\star}_{\pm 0.0}$	$\mathbf{82.8}^{\star}_{\pm 0.2}$	$\mathbf{88.2}^{\star\star}_{\pm 0.2}$	$\mathbf{83.0}^{\star\star}_{\pm 0.9}$

Table 3: F_1 scores of models incorporating different predicate representations and sense distinctions on VerbNet and PropBank SRL on SemLink. SRL + VSD and SRL | VSD are multitask models for SRL and VerbNet classification, with the latter using predicted classes as features for SRL. ALL CLASSES, PREDICTED CLASS, and GOLD CLASS are SRL models using VerbNet class features (the list of all VerbNet classes the predicate belongs to, predicted VerbNet classes, and gold VerbNet classes respectively).

System	WSJ	Brown
ClearWSD	$97.0_{\pm 0}$	$89.3_{\pm 0}$
Baseline	$97.3_{\pm 0.1}$	$90.7_{\pm 0}$
SRL + VSD	$\mathbf{97.7}^{\star\star}_{\pm 0.1}$	$91.3_{\pm 0.4}$
SRL | VSD	$97.6^{\star\star}_{\pm 0.1}$	$91.3_{\pm 0}$

Table 4: VerbNet classification (sense disambiguation) accuracy on SemLink.

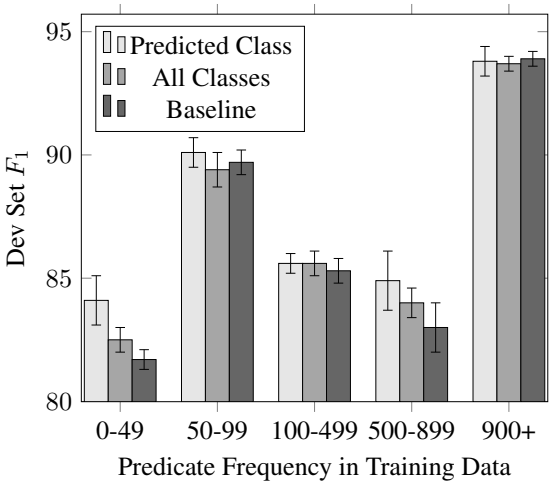

Figure 1: Evaluation by training set predicate frequency on the SemLink development data comparing the impact of VerbNet features.

Focusing on low-frequency predicates, we further divide our evaluation of predicates occurring fewer than 50 times in the training data into 6 bins, one of which is reserved for OOV predicates (Figure 2). From this analysis, we find that VN classes are most impactful on predicates appearing fewer than 10 times in the training data, with a large improvement over the baseline on OOV predicates when applying predicted classes.

8 Conclusions and Future Work

We investigate VerbNet classes as an effective level of abstraction for predicates when performing semantic role labeling. We find that incorporating features based on gold VerbNet classes improves both VerbNet and PropBank SRL, but when predicted classes are used, this effect is only observed for VerbNet. An improvement is also observed without explicit prediction of classes by including a list of all VerbNet classes the target predicate belongs to as features. Breaking down our evaluation into polysemous and monosemous predicates, we find that predicted classes help more on out-of-domain polysemous predicates, while using all valid VerbNet classes helps more on out-of-domain low-frequency predicates. In multi-task learning experiments motivated by the interdependence of VN classification and SRL, we find that joint training improves both tasks when conditioning role labeling on predicted predicates, facilitating VN semantic parsing. In future work, we will investigate alternative approaches incorporating the structure of VerbNet into the parsing of VerbNet semantic representations. Finally, we hope to expand our evaluations to larger, more diverse datasets to further investigate domain transfer.

System	Polysemous WSJ	Brown	Monosemous WSJ	Brown
Baseline	(+0.0) $88.2_{\pm 0.3}$	(+0.0) $81.8_{\pm 0.8}$	(+0.0) $85.9_{\pm 0.3}$	(+0.0) $77.7_{\pm 1.3}$
ALL CLASSES	(+0.4) $\mathbf{88.6}^{\star\star}{}_{\pm 0.2}$	(−0.2) $81.6_{\pm 0.8}$	(+0.2) $86.1^{\star}{}_{\pm 0.4}$	(+2.6) $\mathbf{80.3}^{\star\star}{}_{\pm 0.8}$
PREDICTED CLASS	(+0.3) $88.5^{\star\star}{}_{\pm 0.2}$	(+0.5) $\mathbf{82.3}^{\star}{}_{\pm 0.8}$	(+0.2) $\mathbf{86.1}^{\star\star}{}_{\pm 0.2}$	(+0.9) $78.6^{\star}{}_{\pm 1.3}$

Table 5: Evaluation of contribution of VerbNet features on polysemous vs. monosemous predicates for VerbNet SRL averaged over all models. Average change over the baseline performance is given in parentheses.

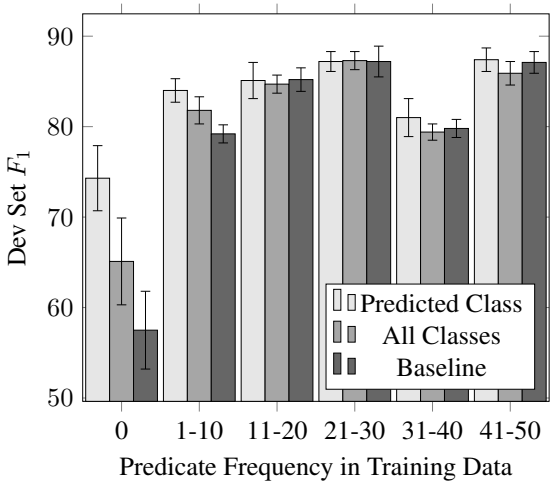

Figure 2: Evaluation by training set predicate frequency similar to Figure 1, but focused on low-frequency predicates. Most improvements are for predicates appearing fewer than 10 times in the training data.

Acknowledgments

We gratefully acknowledge the support of C3 (Cognitively Coherent Human-Computer Communication, subcontracts from UIUC and SIFT), DARPA AIDA Award FA8750-18-2-0016 (RAMFIS), and DTRA HDTRA1-16-1-0002/Project 1553695 (eTASC - Empirical Evidence for a Theoretical Approach to Semantic Components). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any government agency. This work was partially supported by research credits from Google Cloud. Finally, we thank the anonymous IWCS reviewers for their insightful comments and suggestions.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *OSDI*, pages 265–283.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada. Association for Computational Linguistics.

Marzieh Bazrafshan and Daniel Gildea. 2013. Semantic roles for string to tree machine translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 419–423, Sofia, Bulgaria. Association for Computational Linguistics.

Jonathan Berant, Vivek Srikumar, Pei-Chun Chen, Abby Vander Linden, Brittany Harding, Brad Huang, Peter Clark, and Christopher D. Manning. 2014. Modeling biological processes for reading comprehension. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1499–1510, Doha, Qatar. Association for Computational Linguistics.

Claire Bonial, Olga Babko-Malaya, Jinho D Choi, Jena Hwang, and Martha Palmer. 2010. PropBank Annotation Guidelines. Technical report, Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder.

Susan Windisch Brown, Julia Bonn, James Gung, Annie Zaenen, James Pustejovsky, and Martha Palmer. 2019. VerbNet representations: Subevent semantics for transfer verbs. In *Proceedings of the First International Workshop on Designing Meaning Representations*, pages 154–163, Florence, Italy. Association for Computational Linguistics.

Susan Windisch Brown, James Pustejovsky, Annie Zaenen, and Martha Palmer. 2018. Integrating Generative Lexicon event structures into VerbNet. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, pages 152–164, Ann Arbor, Michigan. Association for Computational Linguistics.

Rich Caruana. 1998. Multitask learning. In *Learning to Learn*, pages 95–133. Springer.

Yun-nung Chen, William Yang Wang, and Alexander I Rudnicky. 2013. Unsupervised Induction and Filling of Semantic Slots for Spoken Dialogue Systems Using Frame-Semantic Parsing. *ASRU*, pages 120–125.

Peter Clark, Bhavana Dalvi, and Niket Tandon. 2018. What happened? leveraging verbnet to predict the effects of actions in procedural text. *arXiv:1804.05435*.

Daniel Dahlmeier, Hwee Tou Ng, and Tanja Schultz. 2009. Joint learning of preposition senses and semantic roles of prepositional phrases. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 450–458, Singapore. Association for Computational Linguistics.

Hoa Trang Dang and Martha Palmer. 2005. The role of semantic roles in disambiguating verb senses. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 42–49, Ann Arbor, Michigan. Association for Computational Linguistics.

Dipanjan Das, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, pages 4171–4186.

Kaustubh Dhole and Christopher D. Manning. 2020. Syn-QG: Syntactic and shallow semantic rules for question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 752–765, Online. Association for Computational Linguistics.

David Dowty. 1991. Thematic Proto-Roles and Argument Selection. *Language*, 67(3):547–619.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2785, Florence, Italy. Association for Computational Linguistics.

Nicholas FitzGerald, Oscar Täckström, Kuzman Ganchev, and Dipanjan Das. 2015. Semantic role labeling with neural network factors. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 960–970, Lisbon, Portugal. Association for Computational Linguistics.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Daniel Gildea and Daniel Jurafsky. 2000. Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520, Hong Kong. Association for Computational Linguistics.

James Gung. 2013. Using Relations for Identification and Normalization of Disorders: Team CLEAR in the ShARe/CLEF 2013 eHealth Evaluation Lab. In *CLEF (Working Notes)*.

Kazuma Hashimoto, Caiming Xiong, Yoshimasa Tsuruoka, and Richard Socher. 2017. A joint many-task model: Growing a neural network for multiple NLP tasks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1923–1933, Copenhagen, Denmark. Association for Computational Linguistics.

Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2018. Jointly predicting predicates and arguments in neural semantic role labeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 364–369, Melbourne, Australia. Association for Computational Linguistics.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada. Association for Computational Linguistics.

Karl Moritz Hermann, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1448–1458, Baltimore, Maryland. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Daisuke Kawahara and Martha Palmer. 2014. Single classifier approach for verb sense disambiguation based on generalized features. In *Proceedings of the Ninth International Conference on Language*

Resources and Evaluation (LREC'14), pages 4210–4213, Reykjavik, Iceland. European Language Resources Association (ELRA).

Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*.

Ilia Kuznetsov and Iryna Gurevych. 2020. A matter of framing: The impact of linguistic formalism on probing results. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 171–182, Online. Association for Computational Linguistics.

Zuchao Li, Shexia He, Hai Zhao, Yiqing Zhang, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. 2019. Dependency or Span, End-to-End Uniform Semantic Role Labeling. *AAAI*, 33:6730–6737.

Ding Liu and Daniel Gildea. 2010. Semantic role features for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 716–724, Beijing, China. Coling 2010 Organizing Committee.

Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2017. GUIR at SemEval-2017 task 12: A framework for cross-domain clinical temporal information extraction. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1024–1029, Vancouver, Canada. Association for Computational Linguistics.

Paola Merlo and Lonneke Van Der Plas. 2009. Abstraction and generalisation in semantic role labels: PropBank, VerbNet or both? In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 288–296, Suntec, Singapore. Association for Computational Linguistics.

Paloma Moreda and Manuel Palomar. 2006. The Role of Verb Sense Disambiguation in Semantic Role Labeling. *Lecture Notes in Computer Science Advances in Natural Language Processing*, pages 684–695.

Hiroki Ouchi, Hiroyuki Shindo, and Yuji Matsumoto. 2018. A span selection model for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1630–1642, Brussels, Belgium. Association for Computational Linguistics.

Martha Palmer. 2009. SemLink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*, pages 9–15. Pisa Italy.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

Martha Palmer, James Gung, Claire Bonial, Jinho Choi, Orin Hargraves, Derek Palmer, and Kevin Stowe. 2017. The Pitfalls of Shortcuts: Tales from the Word Sense Tagging Trenches. *Springer series Text, Speech and Language Technology*, Essays in Lexical Semantics and Computational Lexicography - In Honor of Adam Kilgarriff.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Karin Kipper Schuler. 2005. VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. *Dissertation Abstracts International, B: Sciences and Engineering*, 66(6).

Peng Shi and Jimmy Lin. 2019. Simple BERT Models for Relation Extraction and Semantic Role Labeling. *arXiv:1904.05255*.

Kevin Stowe, Sarah Moeller, Laura Michaelis, and Martha Palmer. 2019. Linguistic analysis improves neural metaphor detection. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 362–371, Hong Kong, China. Association for Computational Linguistics.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.

Swabha Swayamdipta, Sam Thomson, Kenton Lee, Luke Zettlemoyer, Chris Dyer, and Noah A. Smith. 2018. Syntactic scaffolds for semantic structures. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3772–3782, Brussels, Belgium. Association for Computational Linguistics.

Gokhan Tur and Dilek Hakkani-Tür. 2005. Semi-Supervised Learning for Spoken Language Understanding Using Semantic Role Labeling. *Language*, pages 232–237.

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

Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2015. Machine comprehension with syntax, frames, and semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 700–706, Beijing, China. Association for Computational Linguistics.

Ronald J. Williams and David Zipser. 1989. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv:1609.08144*.

Qingrong Xia, Zhenghua Li, and Min Zhang. 2019. A syntax-aware multi-task learning framework for Chinese semantic role labeling. 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 5382–5392, Hong Kong, China. Association for Computational Linguistics.

Szu-ting Yi. 2007. *Robust Semantic Role Labeling Using Parsing Variations and Semantic Classes*. Ph.D. thesis, University of Pennsylvania.

Szu-ting Yi, Edward Loper, and Martha Palmer. 2007. Can semantic roles generalize across genres? In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 548–555, Rochester, New York. Association for Computational Linguistics.

Beñat Zapirain, Eneko Agirre, and Lluís Màrquez. 2008. Robustness and generalization of role sets: PropBank vs. VerbNet. In *Proceedings of ACL-08: HLT*, pages 550–558, Columbus, Ohio. Association for Computational Linguistics.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1127–1137, Beijing, China. Association for Computational Linguistics.

Junru Zhou, Zhuosheng Zhang, Hai Zhao, and Shuailiang Zhang. 2020. LIMIT-BERT : Linguistics informed multi-task BERT. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4450–4461, Online. Association for Computational Linguistics.