Verb Knowledge Injection for Multilingual Event Processing

Olga Majewska[1] **Ivan Vulić**[1] **Goran Glavaš**[2] **Edoardo M. Ponti**[1,3] **Anna Korhonen**[1]
[1]Language Technology Lab, TAL, University of Cambridge, UK
[2] Data and Web Science Group, University of Mannheim, Germany
[3] Mila – Quebec AI Institute, Montreal, Canada
[1]{om304,iv250,ep490,alk23}@cam.ac.uk
[2]goran@informatik.uni-mannheim.de

Abstract

Linguistic probing of pretrained Transformer-based language models (LMs) revealed that they encode a range of syntactic and semantic properties of a language. However, they are still prone to fall back on superficial cues and simple heuristics to solve downstream tasks, rather than leverage deeper linguistic information. In this paper, we target a specific facet of linguistic knowledge, the interplay between verb meaning and argument structure. We investigate whether injecting explicit information on verbs' semantic-syntactic behaviour improves the performance of pretrained LMs in event extraction tasks, where accurate verb processing is paramount. Concretely, we impart the verb knowledge from curated lexical resources into dedicated adapter modules (*verb adapters*), allowing it to complement, in downstream tasks, the language knowledge obtained during LM-pretraining. We first demonstrate that injecting verb knowledge leads to performance gains in English event extraction. We then explore the utility of verb adapters for event extraction in other languages: we investigate **1)** zero-shot language transfer with multilingual Transformers and **2)** transfer via (noisy automatic) translation of English verb-based lexical knowledge. Our results show that the benefits of verb knowledge injection indeed extend to other languages, even when relying on noisily translated lexical knowledge.

1 Introduction

Large Transformer-based encoders, pretrained with self-supervised language modeling (LM) objectives, form the backbone of state-of-the-art models for most NLP tasks (Devlin et al., 2019; Yang et al., 2019b; Liu et al., 2019). Recent probes showed that they implicitly extract a non-negligible amount of linguistic knowledge from text corpora in an unsupervised fashion (Hewitt and Manning, 2019; Vulić et al., 2020; Rogers et al., 2020, *inter alia*). In downstream tasks, however, they often rely on spurious correlations and superficial cues (Niven and Kao, 2019) rather than a deep understanding of language meaning (Bender and Koller, 2020), which is detrimental to both generalisation and interpretability (McCoy et al., 2019).

In this work, we focus on a specific facet of linguistic knowledge: reasoning about events.[1] Identifying tokens in the text that mention events and classifying the temporal and causal relations among them is crucial to understand the structure of a story or dialogue (Carlson et al., 2002; Miltsakaki et al., 2004) and to ground a text in real-world facts.

Verbs (with their arguments) are prominently used for expressing events (with their participants). Thus, fine-grained knowledge about verbs, e.g., the syntactic patterns in which they partake and the semantic frames, may help pretrained encoders to achieve a deeper understanding of text and improve their performance in event-oriented downstream tasks. There already exist some expert-curated computational resources that organise verbs into classes based on their syntactic-semantic properties (Jackendoff, 1992; Levin, 1993). In particular, here we consider English VerbNet and FrameNet as rich sources of verb knowledge.

Expanding a line of research on injecting external linguistic knowledge into pretrained LMs (Peters et al., 2019; Levine et al., 2020; Lauscher et al., 2020b), we integrate verb knowledge into the LMs for the first time. We devise a new method to distil verb knowledge into dedicated *adapter* modules (Pfeiffer et al., 2020b), which reduce the risk of (catastrophic) forgetting of and allow seamless modular integration with distributional knowledge.

[1] For instance, in the sentence "*Stately, plump Buck Mulligan came from the stairhead, bearing a bowl of lather (...)*", an event of COMING occurs in the past, with BUCK MULLIGAN as a participant, simultaneously to an event of BEARING with an additional participant, a BOWL.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 6952–6969
August 1–6, 2021. ©2021 Association for Computational Linguistics

We hypothesise that complementing pretrained LMs with verb knowledge should benefit model performance in downstream tasks that involve event extraction and processing. We first put this hypothesis to the test in English monolingual event identification and classification tasks from the TempEval (UzZaman et al., 2013) and ACE (Doddington et al., 2004) datasets. We report modest but consistent improvements in the former, and significant performance boosts in the latter, thus verifying that verb knowledge is indeed paramount for a deeper understanding of events and their structure.

Moreover, expert-curated resources are not available for most of the languages spoken worldwide. Therefore, we also investigate the effectiveness of transferring verb knowledge across languages; in particular, from English to Spanish, Arabic and Chinese. The results demonstrate the success of the transfer techniques, and also shed some light on an important linguistic question: to what extent can verb classes (and predicate–argument structures) be considered cross-lingually universal, rather than varying across languages (Hartmann et al., 2013)?

Overall, our main contributions consist in **1)** mitigating the limitations of pretrained encoders regarding event understanding by supplying external verb knowledge; **2)** proposing a new method to do so in a modular way through verb adapters; **3)** exploring techniques to transfer verb knowledge to resource-poor languages. The performance gains across four diverse languages and several event processing tasks and datasets validate that complementing distributional knowledge with curated verb knowledge is both beneficial and cost-effective.

2 Verb Knowledge for Event Processing

Figure 1 illustrates our framework for injecting verb knowledge from VerbNet or FrameNet and leveraging it in downstream event processing tasks. First, we inject the external verb knowledge, formulated as the so-called *lexical constraints* (Mrkšić et al., 2017; Ponti et al., 2019) (in our case – verb pairs, see §2.1), into a (small) additional set of *adapter parameters* (§2.2) (Houlsby et al., 2019). Second (§2.3), we combine the language knowledge encoded in the original LM parameters and the verb knowledge from *verb adapters* for event processing tasks. To this end, we either *a)* fine-tune both sets of parameters (1. pretrained LM; 2. verb adapters) or *b)* freeze both sets of parameters and insert an additional set of *task-specific adapter parameters*. In both cases, the task-specific training is informed both by the general language knowledge captured in the pretrained LM, and the specialised verb knowledge, captured in the verb adapters.

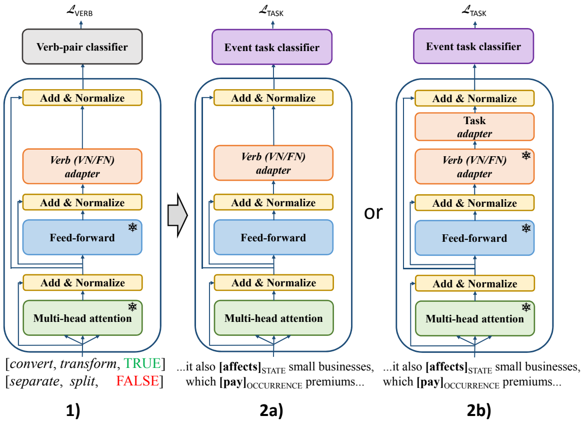

Figure 1: Injecting verb knowledge into a pretrained Transformer-based LM. **1)** Dedicated *verb adapters* trained to recognise pairs of verbs from the same VerbNet (VN) class or FrameNet (FN) frame; **2)** Fine-tuning for an event processing task: **a)** *full fine-tuning* – LM's original parameters and verb adapters both fine-tuned for the task; **b)** *task adapter (TA) fine-tuning* – additional *task adapter* is mounted on top of the *verb adapter* and tuned for the task. For simplicity, we show only a single Transformer layer. Snowflakes denote frozen parameters in the respective training step.

2.1 Sources of Verb Lexical Knowledge

Given the inter-connectedness between verbs' meaning and syntactic behaviour (Levin, 1993; Kipper Schuler, 2005), we assume that refining latent representation spaces with verb knowledge would have a positive effect on event extraction tasks that strongly revolve around verbs. Lexical classes, defined in terms of verbs' shared semantic-syntactic properties, provide a mapping between the verbs' senses and the morpho-syntactic realisation of their arguments (Jackendoff, 1992; Levin, 1993). The potential of verb classifications lies in their predictive power: for any given verb, a set of rich semantic-syntactic properties can be inferred based on its class membership. In this work, we explicitly harness this rich linguistic knowledge to aid pretrained LMs in capturing regularities in the properties of verbs and their arguments.

We select two major English lexical databases – VerbNet (Kipper Schuler, 2005) and FrameNet (Baker et al., 1998) – as sources of verb knowledge at the semantic-syntactic interface, each representing a different lexical framework.

VerbNet (VN) (Kipper Schuler, 2005; Kipper et al., 2006), the largest available verb-focused lexicon, organises verbs into classes based on the overlap in their semantic properties and syntactic behaviour; it builds on the premise that a verb's predicate-argument structure informs its meaning (Levin, 1993). Each entry provides a set of thematic roles and selectional preferences for the verbs' arguments; it also lists the syntactic contexts characteristic for the class members. Its hierarchical classification starts from broader classes and spans several granularity levels where each subclass further refines the semantic-syntactic properties inherited from its parent class.[2] The VN class membership is English-specific, but the underlying verb class construction principles are thought to apply cross-lingually (Jackendoff, 1992; Levin, 1993); its translatability has been indicated in previous work (Vulić et al., 2017; Majewska et al., 2018). The current English VN contains 329 main classes.

FrameNet (FN) (Baker et al., 1998) is more semantically oriented than VN. Grounded in the theory of frame semantics (Fillmore, 1976, 1977, 1982), it organises concepts according to semantic frames, i.e., schematic representations of situations and events, which they evoke, each characterised by a set of typical roles assumed by its participants. The word senses associated with each frame (FN's lexical units) are similar in terms of their semantic content, as well as their typical argument structures. Currently, English FN covers 1,224 frames and its annotations illustrate the typical syntactic realisations of the frame elements. Frames themselves are, however, semantically defined: this means that they may be shared even across languages with different syntactic properties.[3]

2.2 Training Verb Adapters

Training Task and Data Generation. In order to inject external verb knowledge into pretrained LMs, we devise an intermediary training task: we train a dedicated VN-/FN-knowledge adapter (hereafter *VN-Adapter* and *FN-Adapter*). We frame the task as binary word-pair classification: we predict if two verbs belong to the same VN class or FN frame. We extract training instances from FN and VN independently. This allows for a separate analysis of the impact of verb knowledge from each resource.

We generate positive training instances by extracting all unique verb pairings from the set of members of each main VN class/FN frame (e.g., *walk–march*), resulting in 181,882 instances created from VN and 57,335 from FN. We then generate $k = 3$ negative examples per positive example by combining controlled and random sampling. In controlled sampling, we follow prior work on semantic specialisation (Wieting et al., 2015; Glavaš and Vulić, 2018b; Lauscher et al., 2020b). For each positive example $p = (w_1, w_2)$ in the training batch B, we create two negatives $\hat{p}_1 = (\hat{w}_1, w_2)$ and $\hat{p}_2 = (w_1, \hat{w}_2)$; $\hat{w}_1$ is the verb from batch B other than w_1 that is closest to w_2 in terms of their cosine similarity in an auxiliary static word embedding space $X_{aux} \in \mathbb{R}^d$; conversely, $\hat{w}_2$ is the verb from B other than w_2 closest to w_1. We additionally create one negative instance $\hat{p}_3 = (\hat{w}_1,\hat{w}_2)$ by randomly sampling $\hat{w}_1$ and $\hat{w}_2$ from batch B, not considering w_1 and w_2. We ensure that the negatives are not present in the global set of all positive verb pairs.

Similar to Lauscher et al. (2020b), we tokenise each (positive *and* negative) training instance into WordPiece tokens, prepended with sequence start token `[CLS]`, and with `[SEP]` tokens in between the verbs and at the end of the input sequence. We use the representation of the `[CLS]` token $\mathbf{x}_{CLS} \in \mathbb{R}^h$ (with h as the hidden state size of the Transformer) from the last Transformer layer as the latent representation of the verb pair, and feed it to a simple binary classifier:[4] $\hat{\mathbf{y}} = \text{softmax}(\mathbf{x}_{\text{CLS}}\mathbf{W}_{cl} + \mathbf{b}_{cl})$, with $\mathbf{W}_{cl} \in \mathbb{R}^{h\times 2}$ and $\mathbf{b}_{cl} \in \mathbb{R}^2$ as classifier's trainable parameters. We train by minimising the standard cross-entropy loss ($\mathcal{L}_{VERB}$ in Figure 1).

Adapter Architecture. Instead of directly fine-tuning all parameters of the pretrained Transformer, we opt for storing verb knowledge in a separate set of adapter parameters, keeping the verb knowledge

[2] For example, within a top-level class 'free-80', which includes verbs like *liberate*, *discharge*, and *exonerate* which participate in a NP V NP PP.THEME frame (e.g., *It freed him of guilt*), there exists a subset of verbs participating in a syntactic frame NP V NP S_ING ('free-80-1'), within which there exists an even more constrained subset of verbs appearing with prepositional phrases headed specifically by the preposition *from* (e.g., *The scientist purified the water from bacteria*).

[3] For instance, descriptions of transactions will include the same frame elements *Buyer, Seller, Goods, Money* in most languages. Indeed, English FN has inspired similar projects in other languages: e.g., Spanish (Subirats and Sato, 2004), Japanese (Ohara, 2012), and Danish (Bick, 2011).

[4] We also experimented with sentence-level tasks: we fed (a) pairs of sentence examples from VN/FN in a binary classification setup (e.g., *Jackie leads Rose to the store. – Jackie escorts Rose.*); and (b) individual sentences in a multi-class classification setup (predicting the correct VN class/FN frame). These variants, however, led to weaker performance.

separate from the general language knowledge acquired in pretraining. This (1) allows downstream training to flexibly combine the two sources of knowledge, and (2) bypasses the issues with catastrophic forgetting and interference (Hashimoto et al., 2017; de Masson d'Autume et al., 2019).

We adopt the standard efficient adapter architecture of Pfeiffer et al. (2020a,c). In each Transformer layer l, we insert a single adapter ($Adapter_l$) after the feed-forward sub-layer. The adapter itself is a two-layer feed-forward neural network with a residual connection, consisting of a down-projection $\mathbf{D} \in \mathbb{R}^{h \times m}$, a GeLU activation (Hendrycks and Gimpel, 2016), and an up-projection $\mathbf{U} \in \mathbb{R}^{m \times h}$, where h is the hidden size of the Transformer model and m is the dimensionality of the adapter: $Adapter_l(\mathbf{h}_l, \mathbf{r}_l) = \mathbf{U}_l(\text{GeLU}(\mathbf{D}_l(\mathbf{h}_l))) + \mathbf{r}_l$; where $\mathbf{r}_l$ is the residual connection, output of the Transformer's feed-forward layer, and $\mathbf{h}_l$ is the Transformer hidden state, output of the subsequent layer normalisation.

2.3 Downstream Fine-Tuning for Event Tasks

The next step is downstream fine-tuning for event processing tasks. We experiment with (1) token-level event trigger identification and classification and (2) span extraction for event triggers and arguments (a sequence labeling task); see §3. For the former, we mount a classification head – a simple single-layer feed-forward softmax regression classifier – on top of the Transformer augmented with VN-/FN-Adapters. For the latter, we follow the architecture from prior work (M'hamdi et al., 2019; Wang et al., 2019) and add a CRF layer (Lafferty et al., 2001) on top of the sequence of Transformer's outputs (for subword tokens).

For all tasks, we propose and evaluate two different fine-tuning regimes: (1) *full fine-tuning*, where we update both the original Transformer's parameters and VN-/FN-Adapters (see *2a* in Figure 1); and (2) *task-adapter* (**TA**) *fine-tuning*, where we keep both Transformer's original parameters and VN-/FN-Adapters frozen, while stacking a new trainable *task adapter* on top of the VN-/FN-Adapter in each Transformer layer (see *2b* in Figure 1).

2.4 Cross-Lingual Transfer

Creation of curated resources like VN or FN takes years of expert linguistic labour. Consequently, such resources do not exist for a vast majority of languages. Given the inherent cross-lingual nature of verb classes and semantic frames (see §2.1), we investigate the potential for verb knowledge transfer from English to target languages, without any manual target-language adjustments. Massively multilingual LMs, such as multilingual BERT (mBERT) (Devlin et al., 2019) or XLM-R (Conneau et al., 2020) have become the *de facto* standard mechanisms for zero-shot (ZS) cross-lingual transfer. In our first transfer approach: we fine-tune mBERT first on the English verb knowledge, then on English task data, and then simply make task predictions for the target language input.

	VerbNet	FrameNet
English (EN)	181,882	57,335
Spanish (ES)	96,300	36,623
Chinese (ZH)	60,365	21,815
Arabic (AR)	70,278	24,551

Table 1: Number of positive verb pairs in English, and in each target language obtained via VTRANS (§2.4).

The second approach, dubbed VTRANS, is inspired by the work on cross-lingual transfer of semantic specialisation for static word embeddings (Glavaś et al., 2019; Ponti et al., 2019; Wang et al., 2020b). In brief (with full details in Appendix C), starting from a set of positive pairs from English VN/FN, VTRANS involves three steps: (1) *automatic translation* of verbs in each pair into the target language, (2) *filtering* of the noisy target language pairs by means of a transferred relation prediction model trained on the English examples, and (3) *training the verb adapters* injected into the pretrained model, now with the translated and filtered target-language verb pairs. For the monolingual target-language FN-/VN-Adapter training, we follow the protocol used for English, see §2.2.

3 Experimental Setup

Event Processing Tasks and Data. In event processing tasks, systems are tasked with detecting that *something happened*, identifying *what type* of occurrence took place, as well as what *entities* were involved. Verbs typically act as the organisational core of each such event schema, carrying a lot of semantic and structural weight. Therefore, a model's grasp of verbs' properties should have a bearing on final task performance. Based on this assumption, we select event extraction and classification as suitable tasks to profile the methods from §2.

These tasks and the corresponding data are based on the two prominent frameworks for annotating event expressions: TimeML (Pustejovsky et al., 2003, 2005) and the Automatic Content Extraction

(ACE) (Doddington et al., 2004). First, we rely on the TimeML-annotated corpus from *TempEval* tasks (Verhagen et al., 2010; UzZaman et al., 2013), which targets automatic identification of temporal expressions and relations, and events. Second, we use the ACE dataset: it provides annotations for entities, the relations between them, and for events in which they participate in newswire text.[5]

Task 1: Trigger Identification and Classification (TempEval). We frame the first event processing task as a token-level classification problem, predicting whether a token triggers an event and assigning it to one of the following event types: OCCURRENCE (e.g., *died, attacks*), STATE (e.g., *share, assigned*), REPORTING (*e.g., announced, said*), I-ACTION (e.g., *agreed, trying*), I-STATE (e.g., *understands, wants, consider*), ASPECTUAL (e.g., *ending, began*), and PERCEPTION (e.g., *watched, spotted*).[6] We use the TempEval-3 data for English and Spanish (UzZaman et al., 2013), and the TempEval-2 data for Chinese (Verhagen et al., 2010) (see Table 6 in the appendix for exact dataset sizes).

Task 2: Trigger and Argument Identification and Classification (ACE). In this sequence labeling task, we detect and label event triggers and their arguments, with four individually scored subtasks: (i) trigger identification, where we identify the key word conveying the nature of the event, and (ii) trigger classification, where we classify the trigger word into one of the predefined categories; (iii) argument identification, where we predict whether an entity mention is an argument of the event identified in (i), and (iv) argument classification, where the correct role needs to be assigned to the identified event arguments. We use the ACE data available for English, Chinese, and Arabic.[7]

Event extraction as specified in these two frameworks is a challenging, highly context-sensitive problem, where different words (most often verbs) may trigger the same type of event, and conversely, the same word (verb) can evoke different types of event schemata depending on the context. Adopting these tasks for evaluation thus tests whether leveraging fine-grained curated knowledge of verbs' semantic-syntactic behaviour can improve pretrained LMs' reasoning about event-triggering predicates and their arguments.

Model Configurations. For each task, we compare the performance of the underlying "vanilla" BERT-based model (see §2.3) against its variant with an added VN-Adapter or FN-Adapter[8] (see §2.2) in two regimes: (a) full fine-tuning, and (b) task adapter (TA) fine-tuning (see Figure 1). To ensure that any performance gains are not merely due to increased parameter capacity offered by the adapters, we also evaluate a variant where we replace the verb adapter with a randomly initialised adapter of the same size (+*Random*). Additionally, we examine the impact of increasing the capacity of the trainable task adapter by replacing it with a '*Double Task Adapter*' (2TA), i.e., a task adapter with double the number of trainable parameters compared to the base architecture from §2.2. Finally, we compare the VN/FN-Adapter approach with a computationally more expensive alternative method of injecting external verb knowledge, *sequential fine-tuning*, where the full BERT is first fine-tuned on the FN/VN data (as in 2.2) and then on the task (see Appendix D for details).

Training Details: Verb Adapters. We experimented with $k \in \{2, 3, 4\}$ negative examples and the following combinations of controlled (c) and randomly (r) sampled negatives (see §2.2): $k = 2$ $[cc]$, $k = 3$ $[ccr]$, $k = 4$ $[ccrr]$. In our preliminary experiments we found $k = 3$ $[ccr]$ to yield best-performing adapters. The evaluation and analysis presented in §4 are thus based on this setup. Our VN- and FN-Adapters are injected into the BERT Base cased model: the details on adapter training and hyperparameter search are in Appendix B.

Downstream Task Fine-Tuning. In downstream fine-tuning on TempEval, we train for 10 epochs in batches of size 32, with a learning rate $1e-4$ and maximum input sequence length of $T = 128$ WordPiece tokens. For ACE, in light of a greater data sparsity,[9] we search for optimal hyperparameters

[5] We provide more details about the frameworks and their corresponding annotation schemes in Appendix A.

[6] E.g., in the sentence: *"The rules can also affect small businesses, which sometimes pay premiums tied to employees' health status and claims history."*, *affect* and *pay* are event triggers of type STATE and OCCURRENCE, respectively.

[7] The ACE annotations distinguish 34 trigger types (e.g., *Business:Merge-Org*, *Justice:Trial-Hearing*, *Conflict:Attack*) and 35 argument roles. Following previous work (Hsi et al., 2016), we conflate eight time-related argument roles - e.g., 'Time-At-End', 'Time-Before', 'Time-At-Beginning' - into a single 'Time' role in order to alleviate training data sparsity.

[8] We also experimented with inserting both verb adapters simultaneously; however, this resulted in weaker downstream performance than adding each separately, a likely product of the partly redundant, partly conflicting information encoded in these adapters (see §2.1 for comparison of VN and FN).

[9] Most event types ($\approx 70\%$) have fewer than 100 labeled instances, and three have fewer than 10 (Liu et al., 2018).

for each language and evaluation setup from the following grid: learning rate $l \in \{1e-5, 1e-6\}$, epochs $n \in \{3, 5, 10, 25, 50\}$, batch $b \in \{8, 16\}$ (maximum input sequence length $T = 128$).

Transfer Experiments in zero-shot (ZS) setups are based on mBERT, to which we add the VN- or FN-Adapter trained on the English VN/FN data. We train the model on English training data available for each task, and evaluate it on the target-language test set. For the VTRANS approach (§2.4), we use language-specific BERT models available for our target languages, and leverage target-language adapters trained on translated and automatically refined verb pairs. The model, with or without the target-language VN-/FN-Adapter, is trained and evaluated on the training and test data available in the language. We carry out the procedure for three target languages (see Table 1). We use the same negative sampling parameter configuration proven strongest in our English experiments ($k = 3$ [ccr]).

4 Results and Discussion

English Event Processing. Table 2 shows the performance on English Task 1 (TempEval) and Task 2 (ACE). First, we note that the computationally more efficient setup with a dedicated task adapter (TA) yields higher absolute scores compared to full fine-tuning (FFT) on TempEval. When the underlying BERT is frozen along with the added FN-/VN-Adapter, the TA is enforced to encode additional task-specific knowledge into its parameters, beyond what is provided in the verb adapter. This yields two strongest results overall from the +FN/VN setups. On ACE, the primacy of TA-based training is overturned in favour of FFT. Encouragingly, boosts provided by verb adapters are visible regardless of the chosen task fine-tuning regime.

We notice consistent statistically significant[10] improvements in the +VN setup, although the performance of the TA-based setups clearly suffers in argument (ARG) tasks due to decreased trainable parameter capacity. Lack of visible improvements from the Random Adapter supports the interpretation that performance gains indeed stem from the added useful 'non-random' signal in the verb adapters. In addition, we verify how our principal setup with added adapter modules compares to an alternative established approach, sequential fine-tuning (+FN/VN$_{seq}$). In TempEval, we note that fine-tuning all model parameters on VN/FN data allows retrieving more additional verb knowledge beneficial for task performance than adding smaller pre-trained adapters on top of the underlying model. However, FN/VN$_{seq}$ scores are still inferior to the results achieved in the TA-based +FN/VN setup. In ACE, the FN/VN$_{seq}$ results in trigger tasks are weaker than those achieved through the addition of self-contained knowledge adapters, however, they offer additional boosts in argument tasks.

Multilingual Event Processing. Table 3 compares the performance of zero-shot (ZS) transfer and monolingual target training (via VTRANS) on TempEval in Spanish and Chinese. For both, the addition of the FN-Adapter in the TA-based setup boosts ZS transfer. The benefits extend to the FFT setup in Chinese, achieving the top score overall.

In monolingual evaluation, we observe consistent gains from the added transferred knowledge via VTRANS in Spanish. In Chinese performance boosts come from the transferred VN-style class membership information (+VN). This suggests that even the noisily translated verb pairs carry enough useful signal through to the target language. To tease apart the contribution of the language-specific encoders and transferred verb knowledge, we carry out an additional monolingual evaluation substituting the target-language BERT with mBERT, trained on (noisy) target language verb signal (ES-MBERT/ZH-MBERT). Although mBERT scores are lower than monolingual BERTs in absolute terms, the use of the transferred verb knowledge helps reduce the gap between the models, with gains achieved over the baselines in Spanish.[11]

In ACE, the top scores are achieved in the monolingual FFT setting; as with English, keeping the full capacity of BERT parameters unfrozen noticeably helps performance.[12] In Arabic, FN knowledge provides performance boosts across the four tasks and with both the zero-shot (ZS) and monolingual (VTRANS) transfer approaches, whereas the addition of the VN adapter boosts scores in ARG tasks. The usefulness of FN knowledge extends to zero-shot transfer in Chinese, and both adapters benefit the ARG tasks in the monolingual (VTRANS)

[10] We test significance with the Student's t-test with a significance value set at $\alpha = 0.05$ for sets of model F_1 scores.

[11] Due to analogous patterns in relative scores of mBERT and monolingual BERTs in monolingual ACE evaluation, we show the VTRANS mBERT results in ACE in Appendix E.

[12] This is especially the case in ARG tasks, where the TA-based setup fails to achieve meaningful improvements over zero, even with extended training up to 100 epochs. Due to the computational burden of such long training, the results in this setup are limited to trigger tasks (after 50 epochs).

		FFT	+Rand	+FN	+VN	+FN$_{seq}$	+VN$_{seq}$	**TA**	+Rand	+FN	+VN
TempEval	T-IDENT&CLASS	73.6	73.5	73.6	73.6	**74.2**	73.9	74.5	74.4	**75.0**	**75.2**
ACE	T-IDENT	69.3	69.6	**70.8**	**70.3**	70.0	69.8	65.1	65.0	**65.7**	**66.4**
	T-CLASS	65.3	65.5	**66.7**	**66.2**	65.4	65.4	58.0	58.5	**59.5**	**60.2**
	ARG-IDENT	33.8	33.5	34.2	**34.6**	**36.3**	**36.2**	2.1	1.9	2.3	2.5
	ARG-CLASS	31.6	31.6	**32.2**	**32.8**	**34.3**	**33.9**	0.6	0.6	0.8	0.8

Table 2: Results on English TempEval and ACE test sets for full fine-tuning (**FFT**) and the task adapter (**TA**) setup. Provided are average F_1 scores over 10 runs. Statistically significant (paired *t*-test; $p < 0.05$) improvements over both baselines marked in bold; the same labeling is also used in all subsequent tables.

		FFT	+Random	+FN	+VN	**TA**	+Random	+FN	+VN
Spanish	mBERT-zs	37.2	37.2	37.0	36.6	38.0	38.0	**38.6**	36.5
	ES-BERT	77.7	77.1	77.6	77.4	70.0	70.0	**70.7**	**70.6**
	ES-mBERT	73.5	73.6	**74.4**	**74.1**	65.3	65.4	65.8	**66.2**
Chinese	mBERT-zs	49.9	49.9	**50.5**	47.9	49.2	49.5	**50.1**	48.2
	ZH-BERT	82.0	81.6	81.8	81.8	76.2	76.3	75.9	**76.9**
	ZH-mBERT	80.2	80.1	79.9	80.0	71.8	71.8	72.1	71.9

Table 3: Results on Spanish and Chinese TempEval test sets for full fine-tuning (**FFT**) and the task adapter (**TA**) setup, for zero-shot (**zs**) transfer with **mBERT** and monolingual target language evaluation with language-specific BERT (**ES-BERT / ZH-BERT**) or mBERT (**ES-mBERT / ZH-mBERT**), with FN/VN adapters trained on vtrans-translated verb pairs (see §2.4). F_1 scores are averaged over 10 runs.

			FFT	+Random	+FN	+VN	**TA**	+Random	+FN	+VN
Arabic	mBERT-zs	T-IDENT	15.8	13.5	**17.2**	16.3	29.4	30.3	**32.9**	**32.4**
		T-CLASS	14.2	12.2	**16.1**	**15.6**	25.6	26.3	**27.8**	**28.4**
		ARG-IDENT	1.2	0.6	**2.1**	**2.7**	2.0	3.3	3.3	3.6
		ARG-CLASS	0.9	0.4	1.5	**1.9**	1.2	1.6	1.6	1.3
	AR-BERT	T-IDENT	68.8	68.9	**70.2**	68.6	24.0	21.3	**24.6**	23.5
		T-CLASS	63.6	62.8	**64.4**	62.8	22.0	19.5	**23.1**	22.3
		ARG-IDENT	31.7	29.3	**34.0**	**33.4**	–	–	–	–
		ARG-CLASS	28.4	26.7	**30.3**	**29.7**	–	–	–	–
Chinese	mBERT-zs	T-IDENT	36.9	36.7	**42.1**	36.8	47.8	49.4	**55.0**	**55.4**
		T-CLASS	27.9	25.2	**30.9**	**29.8**	38.6	40.1	**43.5**	**44.9**
		ARG-IDENT	4.3	3.1	**5.5**	**6.1**	5.1	6.0	**7.6**	**8.4**
		ARG-CLASS	3.9	2.7	**4.9**	**5.2**	3.5	4.7	**5.7**	**7.1**
	ZH-BERT	T-IDENT	75.5	74.9	74.5	74.9	69.8	69.3	70.0	70.2
		T-CLASS	67.9	68.2	68.0	68.6	58.4	57.5	**59.9**	**60.0**
		ARG-IDENT	27.3	26.1	**29.8**	**28.8**	–	–	–	–
		ARG-CLASS	25.8	25.2	**28.2**	**27.2**	–	–	–	–

Table 4: Results on Arabic and Chinese ACE test sets for full fine-tuning (**FFT**) and the task adapter (**TA**) setup, for zero-shot (zs) transfer with **mBERT** and vtrans transfer approach with language-specific BERT (**AR-BERT / ZH-BERT**) and FN/VN adapters trained on noisily translated verb pairs (§2.4). F_1 scores averaged over 5 runs.

transfer setup. Notably, in zero-shot transfer, we observe that the highest scores are achieved in the task adapter (TA) fine-tuning, where the inclusion of the verb adapters offers additional gains. Overall, however, the argument tasks elude the restricted capacity of the TA-based setup, with very low scores.

Additionally, in Appendix E we show the results with sequential fine-tuning. Similarly to our EN results (Table 2), we observe advantages of using the full capacity of BERT parameters to encode verb knowledge in most setups in TempEval, while the comparison to the adapter-based approach is less clear-cut on ACE. In sum, sequential fine-tuning is a strong verb knowledge injection variant; however, it is computationally more expensive and less portable. The modular and efficient adapter-based approach therefore presents an attractive alternative, while offering competitive task performance. Crucially, the strong results from the sequential setup further corroborate our core finding that external lexical verb information is indeed beneficial for event processing tasks across the board.

Zero-shot Transfer vs Monolingual Training. The results reveal a considerable gap between the performance of zs transfer versus monolingual fine-tuning. The event extraction tasks pose a significant challenge to zero-shot transfer via mBERT; however, mBERT exhibits much more robust performance in the monolingual setup, with available target-language training data for event tasks. In the latter, mBERT trails language-specific BERTs by less than 5 points (Table 3). This is encouraging, given that monolingual pretrained LMs currently exist only for a small set of high-resource languages. For all other languages – should there be language-specific event task data – one can leverage mBERT. Moreover, mBERT's performance is further improved by the inclusion of transferred verb knowledge via VTRANS: in Spanish, where its typological closeness to English renders direct transfer of semantic-syntactic information viable, the addition of VTRANS-based verb adapters yields significant gains both in the FFT and the TA setup.[13] These results confirm the effectiveness of lexical knowledge transfer suggested previously in the work on semantic specialisation of static word vectors (Ponti et al., 2019; Wang et al., 2020b).

Double Task Adapter. Promisingly, we see in Table 5 that the relative performance gains from FN/VN adapters are preserved regardless of the added trainable task adapter capacity. As expected, the increased task adapter size helps argument tasks in ACE, where verb adapters produce additional gains. Overall, this suggests that verb adapters indeed encode additional, non-redundant information beyond what is offered by the pretrained model alone, and boost the dedicated task adapter.

Cleanliness of Verb Knowledge. Despite the promising results with the VTRANS approach, there are still fundamental limitations: (1) noisy translation based on cross-lingual semantic similarity may already break the VerbNet class membership alignment; and (2) the language-specificity of verb classes due to which they cannot be directly ported to another language without adjustments.[14]

The fine-grained class divisions and exact class membership in VN may be too English-specific to allow direct automatic translation. On the contrary, semantically-driven FrameNet lends itself better to cross-lingual transfer: we report higher average gains in cross-lingual setups with the FN-Adapter.

To quickly verify if the noisy direct transfer curbs the usefulness of injected knowledge, we evaluate the injection of *clean* verb knowledge from a small lexical resource available in Spanish: we train an ES FN-Adapter on top of ES-BERT on 2,866 verb pairs derived from its FrameNet (Subirats and Sato, 2004). The results (Appendix E) reveal that, despite having 12 times fewer positive examples for training the verb adapter compared to VTRANS, the 'native' ES FN-Adapter offers gains between +0.2 and +0.4 points over VTRANS, compensating the limited coverage with gold standard accuracy. This suggests that work on optimising and accelerating resource creation merits future research efforts on a par with modeling work.

(a) TempEval		2TA	+FN	+VN
English	EN-BERT	74.5	74.8	74.8
Spanish	mBERT-zs	37.7	**38.3**	37.1
	ES-BERT	73.1	**73.6**	**73.6**
Chinese	mBERT-zs	49.1	**50.1**	48.8
	ZH-BERT	78.1	78.1	**78.6**

	(b) ACE		2TA	+FN	+VN
EN	EN-BERT	T-ID	67.5	**68.1**	**68.9**
		T-CL	61.6	**62.6**	**62.7**
		ARG-ID	6.2	**8.9**	**7.1**
		ARG-CL	3.9	**6.7**	**5.0**
AR	mBERT-zs	T-ID	31.2	**32.6**	31.7
		T-CL	26.3	**27.1**	**29.3**
		ARG-ID	5.9	6.0	**6.9**
		ARG-CL	3.9	4.1	4.3
	AR-BERT	T-ID	40.6	**42.3**	**43.0**
		T-CL	36.9	**38.1**	**39.5**
		ARG-ID	–	–	–
		ARG-CL	–	–	–
ZH	mBERT-zs	T-ID	54.6	**56.3**	**58.1**
		T-CL	45.6	**46.2**	**46.9**
		ARG-ID	9.2	**10.8**	**11.3**
		ARG-CL	8.0	8.5	**9.9**
	ZH-BERT	T-ID	72.3	**73.1**	72.0
		T-CL	59.6	**63.0**	**61.3**
		ARG-ID	2.6	2.8	3.3
		ARG-CL	2.3	2.6	2.9

Table 5: Results on (a) TempEval and (b) ACE for the Double Task Adapter-based approaches (**2TA**).

5 Related Work

Event Extraction. The cost and complexity of event annotation requires robust transfer solutions capable of making fine-grained predictions in the face of data scarcity. Traditional event extraction methods relied on hand-crafted, language-specific features (Ahn, 2006; Gupta and Ji, 2009; Llorens et al., 2010; Hong et al., 2011; Li et al., 2013; Glavaš and Šnajder, 2015) (e.g., POS tags, entity knowledge), which limited their generalisation ability and effectively prevented language transfer.

More recent approaches commonly resorted to word embedding input and neural text encoders such as recurrent nets (Nguyen et al., 2016; Duan et al., 2017; Sha et al., 2018) and convolutional nets (Chen et al., 2015; Nguyen and Grishman, 2015),

[13] We noted analogous positive effects on performance of the more powerful XLM-R Large model (Appendix E).

[14] This is in contrast to the proven cross-lingual portability of synonymy and antonymy relations shown in previous work on semantic specialisation transfer (Mrkšić et al., 2017; Ponti et al., 2019), which rely on semantics alone.

as well as graph neural networks (Nguyen and Grishman, 2018; Yan et al., 2019) and adversarial networks (Hong et al., 2018; Zhang et al., 2019). Most recent empirical advancements in event trigger and argument extraction tasks stem from fine-tuning of LM-pretrained Transformer networks (Yang et al., 2019a; Wang et al., 2019; M'hamdi et al., 2019; Wadden et al., 2019; Liu et al., 2020).

Limited training data nonetheless remains an obstacle, especially when facing previously unseen event types. The alleviation of such data scarcity issues was attempted through data augmentation – automatic data annotation (Chen et al., 2017; Zheng, 2018; Araki and Mitamura, 2018) and bootstrapping for training data generation (Ferguson et al., 2018; Wang et al., 2019). The recent release of the large English event detection dataset MAVEN (Wang et al., 2020c), with annotations of event triggers only, partially remedies for English data scarcity. MAVEN also demonstrates that even the state-of-the-art Transformer models fail to yield satisfying event detection performance in the general domain. The fact that it is unlikely to expect datasets of similar size for other event extraction tasks and especially for other languages only emphasises the need for external event-related knowledge and transfer learning approaches, such as the ones introduced in this work.

Semantic Specialisation. Representation spaces induced through self-supervised objectives from large corpora, be it the word embedding spaces (Mikolov et al., 2013; Bojanowski et al., 2017) or those spanned by LM-pretrained Transformers (Devlin et al., 2019; Liu et al., 2019), encode only distributional knowledge. A large body of work focused on *semantic specialisation* of such distributional spaces by injecting lexico-semantic knowledge from external resources (e.g., WordNet (Fellbaum, 1998), BabelNet (Navigli and Ponzetto, 2010) or ConceptNet (Liu and Singh, 2004)) in the form of lexical constraints (Faruqui et al., 2015; Mrkšić et al., 2017; Glavaš and Vulić, 2018b; Kamath et al., 2019; Vulić et al., 2021).

Joint specialisation models (Yu and Dredze, 2014; Lauscher et al., 2020b; Levine et al., 2020, *inter alia*) train the representation space from scratch on the large corpus, but augment the self-supervised training objective with an additional objective based on external lexical constraints. Lauscher et al. (2020b) add to the Masked LM (MLM) and next sentence prediction (NSP) pretraining objectives of BERT (Devlin et al., 2019) an objective that predicts pairs of (near-)synonyms, aiming to improve word-level semantic similarity in BERT's representation space. In a similar vein, Levine et al. (2020) add the objective that predicts WordNet supersenses. While joint specialisation models allow the external knowledge to shape the representation space from the very beginning of the distributional training, this also means that any change in lexical constraints implies a new, computationally expensive pretraining from scratch.

Retrofitting and post-specialisation methods (Faruqui et al., 2015; Mrkšić et al., 2017; Vulić et al., 2018; Ponti et al., 2018; Glavaš and Vulić, 2019; Lauscher et al., 2020a; Wang et al., 2020a), in contrast, start from a pretrained representation space (word embedding space or a pretrained encoder) and fine-tune it using external lexico-semantic knowledge. Wang et al. (2020a) fine-tune the pre-trained RoBERTa (Liu et al., 2019) with lexical constraints obtained automatically via dependency parsing, whereas Lauscher et al. (2020a) use lexical constraints derived from ConceptNet to inject knowledge into BERT: both adopt adapter-based fine-tuning, storing the external knowledge in a separate set of parameters. Our work adopts a similar adapter-based specialisation approach, however, focusing on event-oriented downstream tasks, and knowledge from VerbNet and FrameNet.

6 Conclusion

We investigated the potential of leveraging knowledge about semantic-syntactic behaviour of verbs to improve the capacity of large pretrained models to reason about events in diverse languages. We proposed an auxiliary pretraining task to inject VerbNet- and FrameNet-based lexical verb knowledge into dedicated verb adapter modules. We demonstrated that state-of-the-art pretrained models still benefit from the gold standard linguistic knowledge stored in lexical resources, even those with limited coverage. Crucially, we showed that the benefits of the knowledge from resource-rich languages can be extended to other, resource-leaner languages through translation-based transfer of verb class/frame membership information.

Acknowledgements. This work is supported by the ERC Consolidator Grant LEXICAL (no 648909) awarded to AK. The work of GG is supported by the Baden-Württemberg Stiftung (Eliteprogramm, AGREE grant).

References

David Ahn. 2006. The stages of event extraction. In *Proceedings of the Workshop on Annotating and Reasoning about Time and Events*, pages 1–8, Sydney, Australia. Association for Computational Linguistics.

Jun Araki and Teruko Mitamura. 2018. Open-domain event detection using distant supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 878–891, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. 2019. Episodic memory in lifelong language learning. In *Advances in Neural Information Processing Systems*, volume 32, pages 13143–13152, Vancouver, Canada. Curran Associates, Inc.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING*, pages 86–90, Montreal, Quebec, Canada.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Eckhard Bick. 2011. A FrameNet for Danish. In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pages 34–41, Riga, Latvia. Northern European Association for Language Technology (NEALT).

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Lynn Carlson, Daniel Marcu, and Mary Ellen Okurowski. 2002. RST Discourse Treebank LDC2002T07. Technical report, Philadelphia: Linguistic Data Consortium.

Yubo Chen, Shulin Liu, Xiang Zhang, Kang Liu, and Jun Zhao. 2017. Automatically labeled data generation for large scale event extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 409–419, Vancouver, Canada. Association for Computational Linguistics.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional 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 167–176, Beijing, China. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Leon R.A. Derczynski. 2017. *Automatically ordering events and times in text*. Springer, Berlin.

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.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 837–840, Lisbon, Portugal. European Language Resources Association (ELRA).

Shaoyang Duan, Ruifang He, and Wenli Zhao. 2017. Exploiting document level information to improve event detection via recurrent neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 352–361, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615, Denver, Colorado. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

James Ferguson, Colin Lockard, Daniel Weld, and Hannaneh Hajishirzi. 2018. Semi-supervised event extraction with paraphrase clusters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 359–364, New Orleans, Louisiana. Association for Computational Linguistics.

Charles J. Fillmore. 1976. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32, New York, New York.

Charles J. Fillmore. 1977. The need for a frame semantics in linguistics. In *Statistical Methods in Linguistics*, pages 5–29. Ed. Hans Karlgren. Scriptor.

Charles J. Fillmore. 1982. Frame semantics. In *Linguistics in the Morning Calm*, pages 111–137. Ed. The Linguistic Society of Korea. Hanshin Publishing Co.

Goran Glavaś, Edoardo Maria Ponti, and Ivan Vulić. 2019. Semantic specialization of distributional word vectors. 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): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

Goran Glavaš and Jan Šnajder. 2015. Construction and evaluation of event graphs. *Natural Language Engineering*, 21(4):607–652.

Goran Glavaš and Ivan Vulić. 2018a. Discriminating between lexico-semantic relations with the specialization tensor model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 181–187, New Orleans, Louisiana. Association for Computational Linguistics.

Goran Glavaš and Ivan Vulić. 2018b. Explicit retrofitting of distributional word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–45, Melbourne, Australia. Association for Computational Linguistics.

Goran Glavaš and Ivan Vulić. 2019. Generalized tuning of distributional word vectors for monolingual and cross-lingual lexical entailment. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4824–4830, Florence, Italy. Association for Computational Linguistics.

Prashant Gupta and Heng Ji. 2009. Predicting unknown time arguments based on cross-event propagation. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 369–372, Suntec, Singapore. Association for Computational Linguistics.

Iren Hartmann, Martin Haspelmath, and Bradley Taylor, editors. 2013. *Valency Patterns Leipzig*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

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.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (GELUs). *CoRR*, abs/1606.08415.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Hong, Jianfeng Zhang, Bin Ma, Jianmin Yao, Guodong Zhou, and Qiaoming Zhu. 2011. Using cross-entity inference to improve event extraction. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1127–1136, Portland, Oregon, USA. Association for Computational Linguistics.

Yu Hong, Wenxuan Zhou, Jingli Zhang, Guodong Zhou, and Qiaoming Zhu. 2018. Self-regulation: Employing a generative adversarial network to improve event detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 515–526, Melbourne, Australia. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799, Long Beach, California, USA. PMLR.

Andrew Hsi, Yiming Yang, Jaime Carbonell, and Ruochen Xu. 2016. Leveraging multilingual training for limited resource event extraction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1201–1210, Osaka, Japan. The COLING 2016 Organizing Committee.

Ray Jackendoff. 1992. *Semantic structures*, volume 18. MIT press.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2984, Brussels, Belgium. Association for Computational Linguistics.

Aishwarya Kamath, Jonas Pfeiffer, Edoardo Maria Ponti, Goran Glavaš, and Ivan Vulić. 2019. Specializing distributional vectors of all words for lexical entailment. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 72–83, Florence, Italy. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representations (ICLR)*, San Diego, CA, USA.

Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. 2006. Extending VerbNet with novel verb classes. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, pages 1027–1032, Genoa, Italy. European Language Resources Association (ELRA).

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvania.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of International Conference on Machine Learning (ICML-2001)*, pages 282–289, Williams College, Williamstown, MA, USA.

Anne Lauscher, Olga Majewska, Leonardo F. R. Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavaš. 2020a. Common sense or world knowledge? Investigating adapter-based knowledge injection into pretrained transformers. In *Proceedings of Deep Learning Inside Out (DeeLIO): The First Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 43–49, Online. Association for Computational Linguistics.

Anne Lauscher, Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2020b. Specializing unsupervised pretraining models for word-level semantic similarity. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1371–1383, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Beth Levin. 1993. *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. 2020. SenseBERT: Driving some sense into BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4656–4667, Online. Association for Computational Linguistics.

Qi Li, Heng Ji, and Liang Huang. 2013. Joint event extraction via structured prediction with global features. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 73–82, Sofia, Bulgaria. Association for Computational Linguistics.

Hugo Liu and Push Singh. 2004. ConceptNet – A practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, and Jun Zhao. 2018. Event detection via gated multilingual attention mechanism. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 4865–4872, New Orleans, Louisiana, USA.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Hector Llorens, Estela Saquete, and Borja Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and semantic roles in TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291, Uppsala, Sweden. Association for Computational Linguistics.

Olga Majewska, Ivan Vulić, Diana McCarthy, Yan Huang, Akira Murakami, Veronika Laippala, and Anna Korhonen. 2018. Investigating the cross-lingual translatability of VerbNet-style classification. *Language resources and evaluation*, 52(3):771–799.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Meryem M'hamdi, Marjorie Freedman, and Jonathan May. 2019. Contextualized cross-lingual event trigger extraction with minimal resources. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 656–665, Hong Kong, China. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26, pages 3111–3119, Lake Tahoe, Nevada, USA. Curran Associates, Inc.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, pages 2237–2240, Lisbon, Portugal. European Language Resources Association (ELRA).

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions* *of the Association for Computational Linguistics*, 5:309–324.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Thien Huu Nguyen, Kyunghyun Cho, and Ralph Grishman. 2016. Joint event extraction via recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 300–309, San Diego, California. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2015. Event detection and domain adaptation with convolutional 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 2: Short Papers)*, pages 365–371, Beijing, China. Association for Computational Linguistics.

Thien Huu Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, volume 18, pages 5900–5907, New Orleans, Louisiana, USA.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics.

Kyoko Ohara. 2012. Semantic annotations in Japanese FrameNet: Comparing frames in Japanese and English. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1559–1562, Istanbul, Turkey. European Language Resources Association (ELRA).

Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. 2019. Knowledge enhanced contextual word representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 43–54, Hong Kong, China.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2020a. AdapterFusion: Non-destructive task composition for transfer learning. *arXiv preprint arXiv:2005.00247*.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020b. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020c. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 282–293, Brussels, Belgium.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Cross-lingual semantic specialization via lexical relation induction. 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 2206–2217, Hong Kong, China. Association for Computational Linguistics.

James Pustejovsky, José M. Castano, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

James Pustejovsky, Robert Ingria, Roser Sauri, José M. Castaño, Jessica Littman, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Inderjeet Mani. 2005. The specification language TimeML. In *The Language of Time: A reader*, pages 545–557. Citeseer.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: what we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Lei Sha, Feng Qian, Baobao Chang, and Zhifang Sui. 2018. Jointly extracting event triggers and arguments by dependency-bridge RNN and tensor-based argument interaction. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pages 5916–5923, New Orleans, Louisiana, USA.

Carlos Subirats and Hiroaki Sato. 2004. Spanish FrameNet and FrameSQL. In *4th International Conference on Language Resources and Evaluation. Workshop on Building Lexical Resources from Semantically Annotated Corpora*, Lisbon, Portugal. Citeseer.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 task 1: TempEval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA. Association for Computational Linguistics.

Marc Verhagen and James Pustejovsky. 2008. Temporal processing with the TARSQI toolkit. In *Coling 2008: Companion volume: Demonstrations*, pages 189–192, Manchester, UK. Coling 2008 Organizing Committee.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 57–62, Uppsala, Sweden. Association for Computational Linguistics.

Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Post-specialisation: Retrofitting vectors of words unseen in lexical resources. 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 516–527, New Orleans, Louisiana. Association for Computational Linguistics.

Ivan Vulić, Nikola Mrkšić, and Anna Korhonen. 2017. Cross-lingual induction and transfer of verb classes based on word vector space specialisation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2546–2558, Copenhagen, Denmark. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Anna Korhonen, and Goran Glavaš. 2021. LexFit: Lexical fine-tuning of pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics.

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

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789, Hong Kong, China. Association for Computational Linguistics.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Cuihong Cao, Daxin Jiang, Ming Zhou, et al. 2020a. K-Adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*.

Shike Wang, Yuchen Fan, Xiangying Luo, and Dong Yu. 2020b. SHIKEBLCU at SemEval-2020 task 2: An external knowledge-enhanced matrix for multilingual and cross-lingual lexical entailment. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 255–262, Barcelona (Online). International Committee for Computational Linguistics.

Xiaozhi Wang, Xu Han, Zhiyuan Liu, Maosong Sun, and Peng Li. 2019. Adversarial training for weakly supervised event detection. 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 998–1008, Minneapolis, Minnesota. Association for Computational Linguistics.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020c. MAVEN: A Massive General Domain Event Detection Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1652–1671, Online. Association for Computational Linguistics.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the Association for Computational Linguistics*, 3:345–358.

Haoran Yan, Xiaolong Jin, Xiangbin Meng, Jiafeng Guo, and Xueqi Cheng. 2019. Event detection with multi-order graph convolution and aggregated attention. 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 5766–5770, Hong Kong, China. Association for Computational Linguistics.

Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. 2019a. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5284–5294, Florence, Italy. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019b. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, pages 5753–5763, Vancouver, Canada.

Mo Yu and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 545–550, Baltimore, Maryland. Association for Computational Linguistics.

Tongtao Zhang, Heng Ji, and Avirup Sil. 2019. Joint entity and event extraction with generative adversarial imitation learning. *Data Intelligence*, 1(2):99–120.

Fanghua Zheng. 2018. A corpus-based multidimensional analysis of linguistic features of truth and deception. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, pages 841–848, Hong Kong. Association for Computational Linguistics.

		Train	Test
TempEval	English	830,005	7,174
	Spanish	51,511	5,466
	Chinese	23,180	5,313
ACE	English	529	40
	Chinese	573	43
	Arabic	356	27

Table 6: Number of tokens (TempEval) and documents (ACE) in the training and test sets.

A Frameworks for Annotating Event Expressions

Two prominent frameworks for annotating event expressions are TimeML (Pustejovsky et al., 2003, 2005) and the Automatic Content Extraction (ACE) (Doddington et al., 2004). TimeML was developed as a rich markup language for annotating event and temporal expressions, addressing the problems of identifying event predicates and anchoring them in time, determining their relative ordering and temporal persistence (i.e., how long the consequences of an event last), as well as tackling contextually underspecified temporal expressions (e.g., *last month, two days ago*). Currently available English corpora annotated based on the TimeML scheme include the TimeBank corpus (Pustejovsky et al., 2003), a human annotated collection of 183 newswire texts (including 7,935 annotated EVENTS, comprising both punctual *occurrences* and *states* which extend over time) and the AQUAINT corpus, with 80 newswire documents grouped by their covered stories, which allows tracing progress of events through time (Derczynski, 2017). Both corpora, supplemented with a large, automatically TimeML-annotated training corpus are used in the TempEval-3 task (Verhagen and Pustejovsky, 2008; UzZaman et al., 2013), which targets automatic identification of temporal expressions, events, and temporal relations.

The ACE dataset provides annotations for entities, the relations between them, and for events in which they participate in newspaper and newswire text. For each event, it identifies its lexical instantiation, i.e., the *trigger*, and its participants, i.e., the *arguments*, and the roles they play in the event. For example, an event type "Conflict:Attack" ("It could swell to as much as $500 billion if we go to war in Iraq."), triggered by the noun "war", involves two arguments, the "Attacker" ("we") and the "Place" ("Iraq"), each of which is annotated with an entity label ("GPE:Nation").

B Adapter Training and Hyperparameter Search

Following Pfeiffer et al. (2020a), we train the adapters for 30 epochs using the Adam algorithm (Kingma and Ba, 2015), a learning rate of $1e-4$ and the adapter reduction factor of 16 (Pfeiffer et al., 2020a), i.e., $d = 48$. Our batch size is 64, comprising 16 positive examples and $3 \times 16 = 48$ negative examples (since $k = 3$).

We experimented with $n \in \{10, 15, 20, 30\}$ training epochs, as well as an early stopping approach using validation loss on a small held-out validation set as the stopping criterion, with a patience argument $p \in \{2, 5\}$; we found the adapters trained for the full 30 epochs to perform most consistently across tasks.

The size of the training batch varies based on the value of k negative examples generated from the starting batch B of positive pairs: e.g., by generating $k = 3$ negative examples for each of 8 positive examples in the starting batch we end up with a training batch of total size $8+3*8 = 32$. We experimented with starting batches of size $B \in \{8, 16\}$ and found the configuration $k = 3$, $B = 16$ to yield the strongest results (reported in this paper).

C VTRANS: Technical Details

First, we automatically translate the verbs by retrieving their nearest neighbour in the target language from the shared cross-lingual embedding space, aligned using the Relaxed Cross-domain Similarity Local Scaling (RCSLS) model of Joulin et al. (2018). Such translation procedure is liable to error due to an imperfect cross-lingual embedding space as well as polysemy and out-of-context word translation. We dwarf these issues in the second step, where we purify the set of noisily translated target language verb pairs by means of a neural lexico-semantic relation prediction model, the Specialization Tensor Model (Glavaš and Vulić, 2018a), here adjusted for binary classification. We train the STM for the same task as verb adapters during verb knowledge injection (§2.2): to distinguish (positive) verb pairs from the same English VN class/FN frame from those from different VN classes/FN frames. In training, the input to STM are static word embeddings of English verbs taken from a shared cross-lingual word embedding space. We then make predictions in the target language by feeding vectors of target language verbs (from noisily translated verb pairs), taken from the same cross-lingual word embedding space, as input for STM. We provide more details on STM training in what follows.

STM Training Details. We train the STM using the sets of English positive examples from each lexical resource (Table 1). Negative examples are generated using controlled sampling (see §2.2), using a $k = 2$ $[cc]$ configuration, ensuring that generated negatives do not constitute positive constraints in the global set. We use the pre-trained 300-dimensional static distributional word vectors computed on Wikipedia data using the FASTTEXT model (Bojanowski et al., 2017), cross-lingually aligned using the RCSLS model of Joulin et al. (2018), to induce the shared cross-lingual embedding space for each source-target language pair. The STM is trained using the Adam optimizer (Kingma and Ba, 2015), a learning rate $l = 1e - 4$, a batch size of 32 (positive and negative) training examples, for a maximum of 10 iterations. We set the values of other training hyperparameters as in Ponti et al. (2019), i.e., the number of specialisation tensor slices $K = 5$ and the size of the specialised vectors $h = 300$.

D Sequential Fine-tuning Details

In the sequential fine-tuning setup, we first train the full cased variant of the BERT-based model on the VN/FN data. We generate negative examples using the strongest performing configuration of sampling parameters: $k = 3$ $[ccr]$. We train the model for 4 epochs using the Adam algorithm (Kingma and Ba, 2015), a learning rate of $2e - 5$ with 1000 warmup steps and a batch size of 64. Next, we fine-tune the VN/FN-pretrained model on the two downstream tasks. For Task 1, we train for 10 epochs in batches of 32 and a learning rate of $1e-4$ and a maximum input sequence $T = 128$. In Task 2, we find an optimal hyperparameter configuration for each language-setup combination from the grid: learning rate $l \in \{1e-5, 1e-6\}$, epochs $n \in \{3, 5, 10, 25, 50\}$, batch size $b \in \{8, 16\}$, with maximum input sequence length of $T = 128$.

E Additional Results

Table 9 presents the results of monolingual evaluation substituting the monolingual target language BERT with the massively multilingual encoder, with or without the FN/VN adapters trained on (noisy) target language verb signal (AR-MBERT/ZH-MBERT). Table 10 provides additional results for Spanish Task 1 (TempEval) using an alternative multilingual encoder, XLM-R (large) (Conneau et al., 2020), as the underlying model (trained with the following hyperparameters: learning rate $l = 2e - 5$, batch size $b \in \{16, 32\}$). Tables 8 and 7 include the results for the sequential fine-tuning setup for Task 1 (TempEval) and Task 2 (ACE), respectively. Table 11 shows the results on Spanish TempEval for different configurations of Spanish BERT with an added Spanish FN-Adapter trained on Spanish FrameNet data.

			+FN$_{seq}$	+VN$_{seq}$
AR	MBERT-ZS	T-IDENT	16.1	15.2
		T-CLASS	15.1	14.1
		ARG-IDENT	1.2	1.1
		ARG-CLASS	1.0	1.0
	AR-BERT	T-IDENT	**70.5**	69.1
		T-CLASS	**65.0**	63.7
		ARG-IDENT	**32.9**	30.2
		ARG-CLASS	**29.5**	27.6
	AR-mBERT	T-IDENT	64.6	65.5
		T-CLASS	55.6	**57.1**
		ARG-IDENT	24.6	23.4
		ARG-CLASS	20.5	19.9
ZH	MBERT-ZS	T-IDENT	**41.6**	**39.9**
		T-CLASS	**29.6**	27.8
		ARG-IDENT	4.6	**7.6**
		ARG-CLASS	4.0	**6.4**
	ZH-BERT	T-IDENT	75.6	75.7
		T-CLASS	**69.0**	68.5
		ARG-IDENT	26.8	26.1
		ARG-CLASS	25.9	25.0
	ZH-mBERT	T-IDENT	72.6	72.6
		T-CLASS	**64.1**	62.2
		ARG-IDENT	**27.0**	24.9
		ARG-CLASS	**25.8**	23.9

Table 7: Results on Arabic and Chinese ACE test sets for the sequential fine-tuning setup for zero-shot (ZS) transfer with **mBERT** and the VTRANS transfer approach with language-specific BERT (**AR-BERT / ZH-BERT**) or mBERT, on noisily translated FN/VN data (§2.4). F_1 scores averaged over 5 runs; significant improvements (paired t-test; $p < 0.05$) over both baselines marked in bold.

		+FN$_{seq}$	+VN$_{seq}$
Spanish	MBERT-ZS	37.3	37.4
	ES-BERT	77.8	77.6
	ES-mBERT	73.3	73.2
Chinese	MBERT-ZS	**51.4**	50.0
	ZH-BERT	82.3	82.2
	ZH-mBERT	80.1	79.1

Table 8: Results on Spanish and Chinese TempEval test sets for monolingual sequential fine-tuning. Significant improvements over the baselines in Table 3 in bold.

			FFT	+Random	+FN	+VN	**TA**	+Random	+FN	+VN
Arabic	AR-MBERT	T-IDENT	64.9	65.2	65.1	65.6	20.7	18.0	**23.2**	19.5
		T-CLASS	56.2	56.5	**57.4**	56.2	14.4	14.0	**16.5**	14.5
		ARG-IDENT	25.4	25.4	**27.2**	24.6	–	–	–	–
		ARG-CLASS	21.3	21.9	**23.0**	19.9	–	–	–	–
Chinese	ZH-MBERT	T-IDENT	74.1	74.4	74.0	73.3	62.2	62.6	**63.8**	62.5
		T-CLASS	62.9	62.9	**64.3**	**63.6**	52.4	52.2	**54.3**	**54.0**
		ARG-IDENT	26.2	26.3	**27.2**	**28.0**	–	–	–	–
		ARG-CLASS	24.8	25.3	**26.2**	**26.4**	–	–	–	–

Table 9: Results on Arabic and Chinese ACE test sets for full monolingual fine-tuning (**FFT**) and the task adapter (**TA**) setup with underlying **mBERT** and VTRANS FN/VN adapters. F_1 scores averaged over 5 runs; significant improvements (paired *t*-test; $p < 0.05$) over both baselines marked in bold.

		FFT	+Random	+FN	+VN	**TA**	+Random	+FN	+VN
Spanish	XLM-R-ZS	42.2	42.3	**42.7**	42.0	40.7	40.3	40.9	40.8
	ES-XLM-R	77.1	77.1	**77.6**	**77.6**	74.8	73.9	74.5	**75.4**

Table 10: Results on Spanish TempEval test sets for full fine-tuning (**FFT**) and the task adapter (**TA**) setup, for zero-shot (**ZS**) transfer and monolingual target language evaluation with XLM-R Large, with FN/VN adapters trained on VTRANS-translated verb pairs (see §2.4). F_1 scores are averaged over 10 runs; significant improvements (paired *t*-test; $p < 0.05$) over both baselines marked in bold.

	FFT+FN$_{ES}$	**TA**+FN$_{ES}$	**2TA**+FN$_{ES}$
ES-BERT	78.0 (+0.4)	70.9 (+0.2)	73.8 (+0.2)

Table 11: Results (F_1 scores) on Spanish TempEval for different configurations of Spanish BERT with added Spanish FN-Adapter (**FN**$_{ES}$), trained on clean Spanish FN constraints. Numbers in brackets indicate relative performance w.r.t. the corresponding setup with FN-Adapter trained on (a larger set of) noisy Spanish constraints obtained through automatic translation of verb pairs from English FN (VTRANS approach).