Where's the Learning in Representation Learning for Compositional Semantics and the Case of Thematic Fit?

Anonymous ACL submission

Abstract

Observing that for certain NLP tasks, such as semantic role prediction or thematic fit estimation, random embeddings perform as well as pretrained embeddings, we explore what settings allow for this and examine where most of the learning is encoded: the word embeddings, the semantic role embeddings, or "the network". We find nuanced answers, depending on the task and its relation to the training objective. We examine these representation learning aspects in multi-task learning, where role prediction and role-filling are supervised tasks, while several thematic fit tasks are outside the models' direct supervision. We observe a non-monotonous relation between some tasks' quality score and the training data size. In order to better understand this observation, we analyze these results using easier, per-verb versions of these tasks.

1 Introduction

We examine to what extent models trained on a simplified semantic role labeling (SRL) task can estimate thematic fit (aka semantic fit), as the training set size grows – and where most of the learning is stored: in the word embeddings, the thematic role embeddings, or elsewhere in the neural net.

A major goal of natural language processing (NLP) is to understand the semantics of language. One traditional NLP task around this is SRL, which labels word spans in a sentence with thematic roles. Consider the sentence "I cut the cake with a knife". We can interpret 'cut' as the action, 'I' as the `Agent` (the performer of the action), 'cake' as the `Theme` of the action (the thing that underwent the action), and 'knife' as the `Instrument` of the action. These words, labeled with roles such as `Agent`, `Theme` and `Instrument`, would be our representation of the event that the sentence conveys. Other sentences with similar meanings, e.g., "the cake was cut with the knife by me", should have the same (or very similar) event representations. In this work, we focus on model training with a simplified version of SRL: each event is represented only by the lemmatized syntactic head of each event argument (including the predicate), and the semantic roles are the simplified PropBank roles (`Arg0, Arg1`, etc.). The reason for this is the current limitations of available evaluation sets for thematic fit: they are all comprised of lemmatized syntactic argument heads as well.

Thematic fit is related to SRL, but separate. This task aims to identify how well a given word or concept fits into a role of an event. Going back to our example sentence, consider these potential replacements for 'knife': scissors, fork and brick. As humans, we understand that while 'knife' is the most typical object for this situation, both 'scissors' and 'fork' could also fit, even if not as naturally. This is because we have a construct of all three objects being instruments for cutting. More so, we know that 'brick' is unlikely to fit given the context of cutting a cake. Since thematic fit datasets are scarce, one challenge in computational linguistics (and computational psycholinguistics) revolves around how machine learning models can learn thematic fit indirectly – perhaps from SRL training. To the best of our knowledge, the state-of-art in this line of work is the residual role-filler averaging model (ResRoFA-MT) proposed by Hong et al. (2018), with an adjusted embeddings representation and training data annotation in Marton and Sayeed (2021).

In this paper, we examine training set size effects on thematic fit tasks – for which the models were not directly optimized – even after reaching a plateau on the simplified SRL task and its complementary task (predicting the head word given the role). **1.** We find surprising training set size interactions with specific evaluation sets and design a modified evaluation metric in order to better understand these interactions. **2.** We also modify the ResRoFA-MT model architecture in various ways to understand what contributes the most to the learning: the pretrained (or random) word embeddings, the thematic role embeddings, or the rest of the network. **3.** In order to be able to train on larger

data, we optimized the code of Hong et al. (2018) and Marton and Sayeed (2021). We release our optimized codebase[1], which trains 6 times faster and includes ablation architectures and a correction to the training data preparation step.

2 Related Work

In event representation models, the main goal is to predict the appropriate word in a sentence given both the role of that supposed word and the surrounding context in the form of word-role pairs. One of the best early neural models was the non-incremental role-filler model (NNRF), by Tilk et al. (2016). This model was based on selectional preferences, or a probability distribution over the candidate words. However, one drawback of this model is that representations of two similarly-worded sentences differing hugely in meaning would closely resemble each other, e.g., "kid watches TV" and "TV watches kid". Another drawback is that the embeddings of the word-role pairs are summed together to represent the sentence, and so the resulting event representation vector does not weight the input vectors differently based on their importance and is not normalized for varying numbers of roles in a sample.

Hong et al. (2018) extend this model in three ways: First, in addition to the word prediction task of NNRF, the task of role prediction given the corresponding word is added, and the two tasks are trained simultaneously (multi-task learning). This model is known as the non-incremental role-filler multitask model (NNRF-MT). Second, they apply the parametric rectified linear unit (PReLU) non-linear function to each word-role embedding, which acts as weights on the composition of embeddings, and subsequently average the embeddings, which normalizes for variable length inputs. This model is called the role-filler averaging model (RoFA-MT). Third, in an effort to tackle the vanishing gradient problem, residual connections between the PReLU output and the averaging input were added together. This third iteration is known as the ResRoFA-MT model. They showed that it performs the best on our thematic fit tasks, and so we use it as our baseline.

Our work differs from Hong et al. (2018) and Marton and Sayeed (2021) in that while they focused more on state-of-the-art performance improvement through new modeling and annotation methods, we aim to understand what controls the learning in such networks.

Previous work suggests a difference between "count" and "predict" models, where "count" models represented lexical semantics in terms of raw or adjusted unsupervised frequencies of correlations between words (such as Local Mutual Information; Baroni and Lenci, 2010) and syntactic or semantic phenomena; "predict" models involve supervised training to achieve their representations, e.g., neural models. Baroni et al. (2014) do a systematic exploration of tasks vs. state-of-the-art count and predict models and found that predict models were overall superior, including for thematic fit tasks. More recently, Lenci et al. (2022) demonstrate that predict-models are not reliably superior to count-models, but depend on the task and the way the models are trained. They also show that even recent contextual models such as BERT are not necessarily better for out-of-context tasks than well-tuned static representations, predict or otherwise.

3 Datasets

We use the Rollenwechsel-English, Version 2 (RW-Eng v2) corpus (Marton and Sayeed, 2021) as the training set for all our experiments. This corpus is sentence-segmented, annotated with morphological analyses, syntactic parses, and syntax-independent PropBank-based semantic role labeling (SRL). The syntactic head word of each semantic argument is determined by using several heuristics to match the parses to the semantic argument spans. Note that a sentence may have multiple predicates (typically verbs) and therefore multiple semantic frames (sometimes called "events"), each with its own semantic arguments, whose span may overlap the argument span of other frames in the sentence.

The first version of this corpus contained NLTK lemmas, MaltParser parses, parts-of-speech (POS) tags, and SENNA SRL tags (Bird, 2006; Nivre et al., 2006; Collobert and Weston, 2007). The second version added layers from more modern taggers: Morfette lemmas, spaCy syntactic parses and POS tags, and LSGN SRL tags (Chrupala, 2011; Honnibal and Johnson, 2015; He et al., 2018). In our experiments here we use the lemmas of the semantic arguments' head words in v2.

The sentences themselves are taken from both the ukWaC (Ferraresi et al., 2008) and the British National Corpus (BNC). This corpus contains 78M sentences across 2.3M documents. This in-

[1] Anonymized

cludes 210M verbal predicates with 700M associated role-fillers. We use the same training, validation, and test split as Hong et al. (2018). That is, we have 99.2% (201.5M samples) in the full training set, 0.4% in validation, and 0.4% in testing. We run our training experiments on different subsets of the training data, ranging from 1% up to the full dataset. We cap our vocabulary size at the 50,000 most common words in that specific subset.

We used the following psycholinguistic test sets:

Padó (Padó et al., 2006) 414 verb-argument pairs and the associated judgement scores. These were constructed from 18 verbs that are present in both FrameNet and PropBank. For each verb, the three most frequent subjects and objects from each of the underlying corpora were selected. That yielded six arguments per verb per corpus, with some overlap between corpora. For each verb-argument pair, a judgement was collected online with an average of 21 ratings per item for the argument in subject and object role. The rating was collected on a Likert scale of 1-7 with the question "How common is it for [subject] to [verb]?" or "How common is it for [object] to be [verbed]?"

McRae (McRae et al., 1998) 1444 pairs of verb-argument pairs in a similar format to Padó. These were created using a similar rating question as the Padó dataset, but is a compilation of ratings collected over several studies with considerable overlap and heterogeneous selection criteria.

Ferretti-Instruments and Ferretti-Locations (Ferretti et al., 2001) 274 predicate-location pairs and 248 predicate-instrument pairs. Based on the McRae dataset (Psychological norms).

GDS (Greenberg et al., 2015) 720 predicate-object pairs and their ratings. Only objects (no subjects), matched for high and low polysemy and frequency, well fitting vs. poorly fitting. Greenberg and McRae overlap by about a third, but the human scores are obtained from new surveys.

Bicknell (Bicknell et al., 2010) 64 cases. Congruent vs incongruent `Patient` in an `Agent`-Verb-`Patient` paradigm. Hand crafted, not corpus-based, designed for event-related potentials-based neurolinguistic experiments.

4 Modeling and Methodology

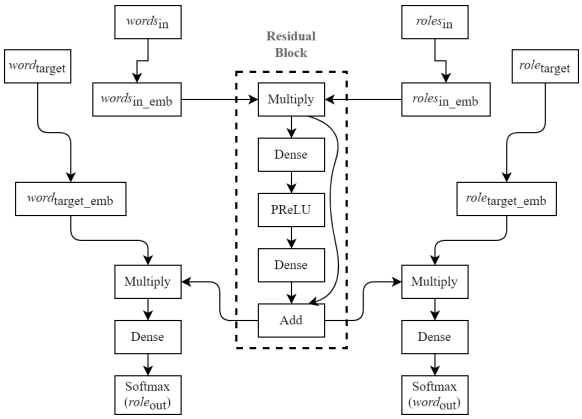
In this setup, an input event is represented as role-word pairs, where the role is one of the following PropBank (Palmer et al., 2005) roles: `Arg0`, `Arg1`, `ArgM-Mnr`, `ArgM-Loc`, `ArgM-Tmp`, and the predicate. The word is the argument's syntactic head's lemma. Both the role and the head word are taken from RW-Eng v2. All prior works with the ResRoFA-MT model use two random word embedding sets (one for input words and one for the target word) and similarly two role embedding sets. See Figure 1a.

Our implementation differs in these key aspects:

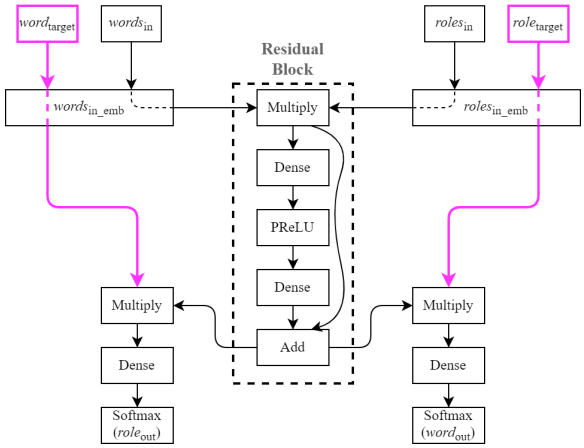
- **Modified model architecture** - Using a single word embeddings set, shared between the target and input words, and similarly a single role embeddings set (Figure 1b). In our experiments, we find the non-shared, redundant embedding layers do not affect the performance while adding (vocab size 50,000 $\times$ word embedding size 300) 15,000,000 learnable parameters in the model.
- **Changes in Batching** - With previous implementations, one "epoch" only resulted in about a third of the data being traversed. The next epoch would start on the second third and so on. Now, we set the data preprocessing so that "one epoch = one pass through all the training data". Additionally, the data is preprocessed during the training of each batch, so no time is lost during training in waiting for the next batch of data to be preprocessed.
- **Missing and unknown words handling** - Following Marton and Sayeed (2021) but unlike Hong et al. (2018), we represent out-of-vocabulary (OOV) words separately from missing words (empty slots in an event).

5 Experiments and Discussion

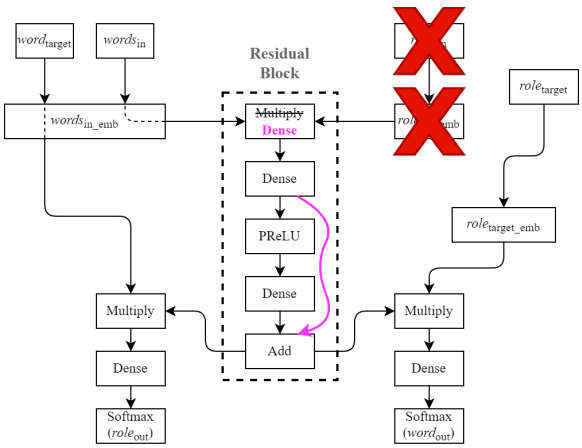
It has been repeatedly observed that in some settings, random word embeddings perform as well as pretrained ones, or very nearly, including in our baselines (Tilk et al., 2016; Hong et al., 2018; Marton and Sayeed, 2021). We design experiments to answer the following questions: **Q1.** Why is this so in our compositional semantics and psycholinguistic tasks? **Q2.** For such semantic tasks and architecture, where is the learning encoded? Is it in the word embeddings, role embeddings or "the network"? **Q3.** Training set size effect: is more data better for this indirect setting and tasks?



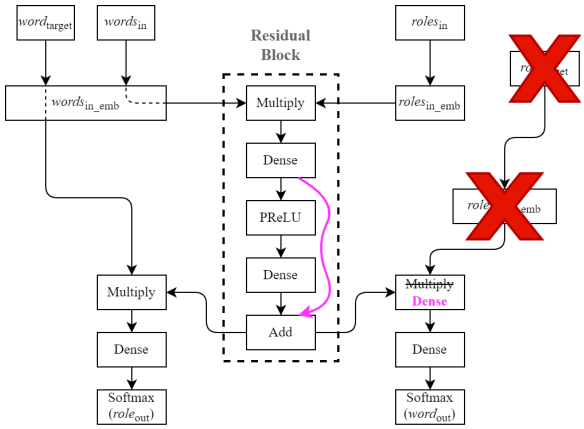
(a) Original ResRoFA-MT architecture



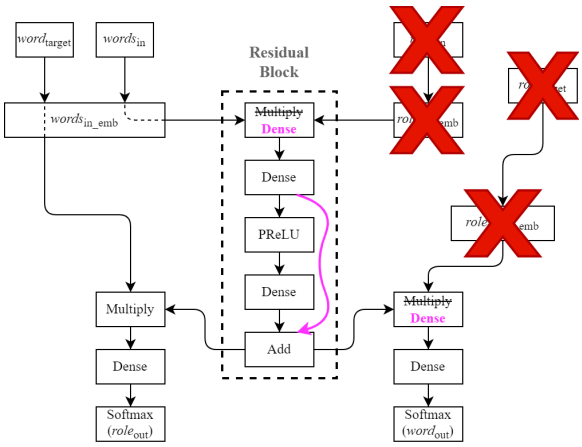
(b) Shared Embedding Layer



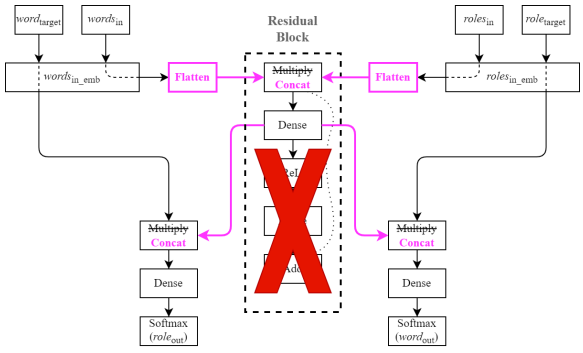
(c) **NIR** network



(d) **NTR** network



(e) **NR** network



(f) Simple network

Figure 1: The model architectures for our experiments

5.1 Objective and Evaluation

We train a feed-forward network in a multi-task learning setting to optimize word and role prediction accuracy. For target word prediction we give the prediction layer a context vector formed as a multiplication of the input word-role pairs and the target role. Similarly, for target role prediction we feed the same context vector along with the target word, following the ResRoFA-MT architecture (Hong et al., 2018) (Figure 1a). Since the network initialization is random, we perform 5 runs of each experiment and report the mean with a 95% con-

fidence interval. Following Hong et al. (2018); Marton and Sayeed (2021), we test each model on the psycholinguistic datasets (Section 3), for which the models were not directly optimized. The idea behind using the latter test battery is that the model, even though trained on (simplified) SRL and word prediction (aka role-filling) tasks, is expected to be able to make indirect generalizations about predicate–argument fit level from the training data and the related objectives. These psycholinguistic tasks are evaluated with Spearman's rank correlation between the sorted human scores and the sorted model scores, except for Bicknell, for which we take accuracy of predicting which argument in each `Patient` role-filler pair is (more) congruent (Lenci, 2011).

5.2 Shared Embedding Layer

We modify the network to use a single embeddings set shared between the input words and target word, by using a single index-to-embedding mapping layer – and similarly a shared embedding-mapping layer for the input roles and target role (Figure 1b). This change results in 2x the training speed (Section 4) without degradation in performance (see first two rows in Tables 1 and 2). Therefore we use the faster shared architecture for the rest of the experiments. We train all models (until Section 5.6) on a uniformly sampled 1% subset, which is large enough to get indicative results while saving time and cost in experimentation. For comparison of our results to previous work, see Section 5.6.

5.3 Random vs. Pre-trained Embeddings

Hong et al. (2018) used random Glorot uniform to initialize the word embeddings. Private communication with the authors confirmed random embeddings do as well as pretrained ones for these tasks. We replicate this finding, comparing random word embeddings to pretrained GloVe embeddings (Pennington et al., 2014), both of size 300. See the third row in the top part of Tables 1 and 2.

(**Q1**) Why is this so? We note that during training, embeddings get updated. To check if this update is responsible for bridging the gap between zero knowledge (random embeddings) and much knowledge (compressed in the pre-trained GloVe), we freeze the word embedding layer and rerun the experiments (see the middle part in the same two tables). Contrary to our previous experiment, we find fixed GloVe embeddings do much better than fixed random embeddings on all our tasks. We also see tuning helps the network converge much faster (from 25 epochs down to 11-15).

We conclude that indeed much of the learning is captured in the word embeddings. Tuning them even on only 1% of our training data bridges the knowledge gap from the pre-trained embeddings almost completely (with possible exceptions on Ferretti and Bicknell). But we note that although lower, the fixed embeddings results are not near-random. This leads us to **Q2.**: Where else is learning done, and to what extent?

5.4 Role Contribution

We now turn to role ablation tests. First we take away the input roles from the context embeddings and call this the no-input-roles network **NIR** (see Figure 1c and the third part of Tables 1 and 2). We do not see significant drops in most of the tasks except role prediction, which we expect by construction. Note that when predicting the target word, the NIR network still receives the target role information,which, together with at least the predicate, is likely often sufficient information for prediction.

We find it surprising that input role ablation barely affects performance on the psycholinguistic tasks. Why is that? One possibility is that the input role contribution is negligible. But another possibility is that in NIR, all (or almost all) the role information had to be 'crammed' into the target role embeddings. To tease these apart, we next take away the target role from the penultimate layer of the network, but leave the input roles intact. We call this no-target-role network **NTR** (see Figure 1d and the row after NIR in the same tables). Now the role accuracy goes back to the base level (as expected by construction), but word accuracy as well as performance on the psycholinguistic tasks drop. We conclude that target role carries more crucial information than input roles for our psycholinguistic tasks, and that role information 'cramming', if it happens in NIR, does not happen in the other direction (NTR).

Finally, for completeness, we remove all role information from the network. We call this no-role network **NR** (see Figure 1e and same tables). This results in a drastic drop in word accuracy as well as the psycholinguistic tasks. This is an an interesting finding which supports previous knowledge about the importance of roles in multi-task learning setting while at the same time defies the importance of roles in the context vector (the output of the resid-

Embedding	Shared?	Tuned?	Role?	Role Accuracy	Word Accuracy	Epochs*
Random	N	Y	Y	$.9655 \pm .0014$	$.1363 \pm .0020$	11(6)
Random	**Y**	Y	Y	$.9671 \pm .0003$	$.1372 \pm .0022$	11(6)
GloVe	Y	Y	Y	$.9669 \pm .0003$	$.1374 \pm .0005$	15(10)
Random	Y	**N**	Y	$.6609 \pm .0046$	$.1208 \pm .0012$	25(20)
GloVe	Y	N	Y	$.9510 \pm .0011$	$.1291 \pm .0006$	25(20)
GloVe	Y	Y	**NIR**†	$.9036 \pm .0013$	$.1348 \pm .0019$	11(6)
GloVe	Y	Y	**NTR**†	$.9677 \pm .0006$	$.1230 \pm .0017$	12(7)
GloVe	Y	Y	**NR**†	$.9007 \pm .0021$	$.1078 \pm .0010$	8(3)
RAND Network‡	Y	Y	Y	$.1530 \pm .0716$	$.0000 \pm .0000$	-
Simpler Network+	Y	N	Y	$.9987 \pm .0005$	$.1208 \pm .0020$	6(1)

Table 1: Word and Role accuracy on 1% training data.

† **NIR**=No input role (in context); **NTR**=No target role (in prediction); **NR**=No role
‡ Network with no training that uses previously fine tuned word/role embeddings as input
+ Simpler Feed forward Network with previously fine tuned word/role embeddings as input
* Epochs in parentheses: the epoch of the effective model (best model before early stopping after patience limit)

Embed.	Shrd	Tuned	Role	Padó	McRae	GDS	Ferretti-Loc	Ferretti-Instr	Bicknell
Random	N	Y	Y	$.5474 \pm .0345$	$.3231 \pm .0236$	$.4485 \pm .0314$	$.2611 \pm .0036$	$.2282 \pm .0623$	$.5260 \pm .1185$
Random	**Y**	Y	Y	$.5280 \pm .0274$	$.3384 \pm .0174$	$.4388 \pm .0206$	$.2532 \pm .1421$	$.2266 \pm .0391$	$.5000 \pm .0673$
GloVe	Y	Y	Y	$.5316 \pm .0320$	$.3280 \pm .0177$	$.4534 \pm .0209$	$.2851 \pm .0301$	$.2895 \pm .0258$	$.5438 \pm .0370$
Random	Y	**N**	Y	$.4396 \pm .0344$	$.2838 \pm .0109$	$.2841 \pm .0246$	$.1767 \pm .0273$	$.2086 \pm .0322$	$.4781 \pm .0450$
GloVe	Y	N	Y	$.4941 \pm .0247$	$.3090 \pm .0254$	$.4349 \pm .0229$	$.3011 \pm .0301$	$.3439 \pm .0421$	$.5563 \pm .0490$
GloVe	Y	Y	**NIR**	$.5079 \pm .0587$	$.3205 \pm .0580$	$.4217 \pm .0472$	$.3054 \pm .0791$	$.2543 \pm .0796$	$.6042 \pm .0896$
GloVe	Y	Y	**NTR**	$.2400 \pm .0294$	$.0937 \pm .0258$	$.3845 \pm .0083$	$.3071 \pm .0017$	$.2621 \pm .0531$	$.5469 \pm .0388$
GloVe	Y	Y	**NR**	$.2496 \pm .1088$	$.1139 \pm .0150$	$.3385 \pm .0363$	$.2955 \pm .1243$	$.2668 \pm .0375$	$.5885 \pm .0448$
RAND	Y	Y	Y	$-.0001 \pm .1090$	$.0109 \pm .1604$	$.0365 \pm .0784$	$.0165 \pm .1048$	$-.0346 \pm .0785$	$.4531 \pm .1027$
Simpler	Y	N	Y	$.3271 \pm .0555$	$.2175 \pm .0294$	$.3356 \pm .0345$	$.1055 \pm .0259$	$.0459 \pm .1239$	$.5365 \pm .0593$

Table 2: Thematic Fit tests on 1% training data (same models as in Table 1)

ual block in Figure 1). Next, we turn to learn more about the impact this vector and the block it is in.

5.5 "It's the Network!"... Or is it?

In order to see how much the particular ResRoFA-MT model architecture (aka "the network") contributes in our tasks, we first use the fine-tuned GloVe embedding from a previously trained base model (third row in Table 1) and assign the rest of the network random weights ("RAND Network" in Tables 1 and 2). To ensure the random weights are similar in size to the trained weights, we calculate the mean and standard deviation for each layer separately and assign that layer random weights using a Gaussian distribution with the same parameters. We see this new model does very poorly, near random prediction. This could be due to the learned representation in the network weights that were ablated here but also due to incompatibility of the non-trained random network weights with the very informative word embeddings.

Therefore next we replace the complex middle "residual block" with a plain dense projection layer but let this ("Simpler Network" (Figure 1f, Tables 1 and 2) learn during training. In training here we use the fine-tuned word (and role) embeddings from our base model. Curiously, we see a large jump in role accuracy, but a drop in word accuracy as well as psycholinguistic tasks other than Bicknell's. We can only speculate as to why the latter task is an outlier here. It involves comparing the plausibility of two two-participant events with one participant changed. A simpler network may have an easier time representing binary distinctions within a pair of simple events, as opposed to predicting fine-grained scores of more complex inter-relationships, evaluated by the use of Spearman's ρ in the other datasets. It may even be able to rely on general collocation statistics here, regardless of roles, but we leave this for future work. Note that here, we still do multi-task prediction as before, but in a much simpler network.

This, along with the role ablation experiments, suggest that while the potential incompatibility of the non-trained random network weights with the word embeddings may account for some of the

drop in performance, the context vector formation through multiplication and likely also the improvements implemented in our base model have a large impact on the representation learning as tested on the thematic fit tasks (although not the same impact on role/word prediction).

We see again that there is no clear correlation between the increase in directly optimized for role/word prediction, and the performance on the psycholinguistic tasks for which the models were not directly optimized.

To recap, it seems that the answer to **Q2** is nuanced: Padó and McRae are most sensitive to ablated roles; GDS, and perhaps also Bicknell, to non-tuned random word embeddings; Ferretti to ablated (simplifiled) networks; and all are sensitive to RAND Networks, but Bicknell is surprisingly robust even there.

5.6 Training Data Size Effect

Often in machine learning and NLP, models learn better with more data. However, there are typically diminishing returns. To test the effect of training data size, we use our shared layer network with tuned GloVe embeddings (as in row 3 in Table 1) on uniformly sampled 1%, 10%, 20% 40% and 100% of the training dataset. See Table 3 and Table 4.

Sys	Role Accuracy	Word Accuracy	Epochs
B1[†]	.9470	-	-
B2[‡]	.9715 ± .0010	.1541 ± .0045	-
20%M[+]	.9707 ± .0002	.1450 ± .0004	-
0.1%	.9446 ± .0015	.0994 ± .0024	12(7)
1%	.9669 ± .0003	.1374 ± .0005	15(10)
10%	.9701 ± .0002	.1443 ± .0006	13(10)
20%	.9703 ± .0004	.1445 ± .0009	9(6)
40%	.9704 ± .0007	.1442 ± .0011	9(6)
100%[2]	.9708 ± .0006	.1444 ± .0019	7(4)

Table 3: Comparison of performance with GloVe (tuned) with varying training set sizes (Sys)

[†] Hong et al. (2018) 20%?
[‡] Marton and Sayeed (2021) 20%
[+] The average of max value in each trial for fair comparison with benchmarks B1,B2

First, in order to compare fairly with previous work, we report the average of the *maximum* value in each training trial on 20% of the data. (Recall that our 20% of the data is a larger training set than our baselines' 20% due to improvements in our batcher). Our role accuracy is better than Hong et al. (2018) and similar to Marton and Sayeed (2021). Our word accuracy is a bit lower than the latter. On the indirectly supervised thematic fit tasks, our results are better on Padó, similar on McRae, but lower for the rest. We suspect that in the previous work authors reported the *best* of all the epochs from all trials, which can explain why the previously reported scores are higher than our results; but we could not verify that.

In order to better understand the effect of training set size (**Q3**), we use next what we believe to be more realistic numbers: the average of the last saved model in each run (best model per our validation set) in each training subset size.

We see incremental improvements from 0.1% to 1% to the 10% dataset across all our evaluation tasks; however, contrary to our null hypothesis, we see diminishing returns or no gains in role and word prediction when using 20% or more of the training set. In most of the psycholinguistic tasks (Table 4), results plateau at 10% or 20% with the notable exception of Padó and McRae, where we see a negative trend beyond 20%. Why is it so, and only for these two tasks, with mainly Padó? The Padó dataset is constructed from high-frequency fillers. It behaves differently from the other datasets and gets a high maximum average score on the 20% subset probably because there is more training data available for high-frequency fillers, compared to the other datasets, including McRae. Considering the small samples in these test sets, they might quickly become victims of not only high variance, but also of overfitting, that is to say, the models may specialize on the corpus distribution, increasingly with training set size. This distribution is likely to be different from the WSJ distribution, from which Padó dataset is drawn (but see also Section 5.7).

How do word/role prediction and thematic fit tasks relate to each other? We leave this question for future research, but our hypothesis is that psycholinguistic meaning of natural language is grounded in interaction with other modalities (e.g., actions, vision, audio), which a model cannot learn just from more textual training data.

This leads potentially to a much bigger question: how much can a neural model learn natural language by just being trained on very large corpora or billions of parameters, and where is the saturation point? Furthermore, we see role information is important to our psycholinguistic tasks; how much does the role definition and granularity (e.g., PropBank or FrameNet), or the role set size, matter for these tasks? Possibly, with a richer roleset, we may see more alignment between word / role prediction and the psycholinguistic tasks. Perhaps PropBank

System	Padó	McRae	GDS	Ferretti-Loc	Ferretti-Instr	Bicknell
B1	.5300	.4250	.6080	.4630	.4770	.7450
B2	$.5363 \pm .0035$	$.4322 \pm .0232$	-	-	-	-
20%M	$.5855 \pm .0101$	$.4338 \pm .0181$	$.5495 \pm .0220$	$.3539 \pm .0239$	$.4255 \pm .0210$	$.6094 \pm .0000$
0.1%	$.2992 \pm .0441$	$.1856 \pm .0157$	$.1699 \pm .0180$	$.0891 \pm .0306$	$.0367 \pm .0203$	$.4906 \pm .0402$
1%	$.5316 \pm .0320$	$.3280 \pm .0177$	$.4534 \pm .0209$	$.2851 \pm .0301$	$.2895 \pm .0258$	$.5438 \pm .0370$
10%	$.5572 \pm .0247$	$.3993 \pm .0137$	$.5409 \pm .0150$	$.3410 \pm .0358$	$.3765 \pm .0320$	$.5906 \pm .0320$
20%	$.5241 \pm .0558$	$.3708 \pm .1182^{\dagger}$	$.5245 \pm .0148$	$.3191 \pm .0312$	$.3853 \pm .0454$	$.5813 \pm .0210$
40%	$.3662 \pm .1355$	$.3831 \pm .0276$	$.5467 \pm .0183$	$.3331 \pm .0215$	$.3660 \pm .0284$	$.5750 \pm .0460$
100%‡	$.3375 \pm .7293$	$.3733 \pm .5203$	$.5338 \pm .1328$	$.2736 \pm .7846$	$.3416 \pm .3297$	$.6094 \pm .1985$

Table 4: Thematic Fit with GloVe tuned (same models as in Table 3)

† 1 trial had an outlier score .2026
‡ All experiments had 5 runs per training subset, except for the 100% with only 2 runs, due to compute resource limitation.

roles are too coarse-grained to allow for an analysis of how a role-prediction task relates to a thematic fit task, which involves the fine-grained ranking (via Spearman's ρ) of event plausiblities derived from the underlying semantic characteristics of the nouns and verbs involved. If so, understanding how performance on a role-prediction task relates to thematic fit judgements may not be possible without a finer-grained inventory of semantic characteristics, such as Dowtyan proto-roles (Dowty, 1991).

5.7 Global and Local Correlation

We evaluate both Padó and McRae by computing Spearman's rank correlation between the sorted list of model's probability scores and the sorted list of averaged human scores, for each dataset. Why do Padó and McRae deteriorate with increasing training data size? To test if this is due to fluctuation of model scores for unrelated but near-in-score verb-noun pairs, we averaged correlations for "local" subsets, grouped by verb. This should be an easier task, since some of the globally close competition is not present in each by-verb subset. Indeed, we see high jumps of 5-8% for the "local" correlation scores in the larger subsets (40% and 100%). But in the smaller subsets we see changes of 2-3% up or down. Moreover, the trend of lower correlation with larger training sets remained. We leave it to future work to dig further into why Padó and McRae show such an anomaly.

6 Conclusions and Future Work

In this work, we explored why random word embeddings counter-intuitively perform as well as pretrained word embeddings on certain compositional semantic tasks (some being outside the models' explicit objective), where the learning is actually stored (teasing apart the word embeddings, role embeddings, and the rest of the network), and how training set size affects performance on these tasks. We found out that tuning (or further tuning) the word embeddings helps and can bridge the gap between random and pretrained embeddings. Moreover, our tuned embedding space is different from pretrained embeddings like GloVe. We saw that the target role is more important than the input roles on our tasks. Furthermore, our experiments suggested that much of the learning happens also in the rest of the network outside word and role embedding layers. No single factor (word / role embeddings or the network) is most important for all tasks.

Training set size had a surprising negative effect on Padó and McRae beyond 20% of the training data. We attempted explaining this with an alternative evaluation method, but this remains to be explained further.

We release our code, including our preferred network architecture – a modified version of ResRoFA-MT with shared embedding layers.

One avenue in which we want to invest is to better understand the complex relationship between word/role accuracy and our psycholinguistic tasks. While our initial hypothesis was that training the network to minimize loss on word/role prediction would also optimize performance on all our tasks, this did not always hold. We suspect that the groundedness is the missing link for (artificially and naturally) learning psycholinguistic tasks, and therefore adding grounding seems promising to us.

Another future avenue for us is investigating the high variability in performance on psycholinguistic tasks, compared to the fairly stable results on the directly optimized-for word and role prediction tasks.

Ethical Considerations

Our work uses RW-Eng v2 (Marton and Sayeed, 2021), which in turn uses two corpora: ukWaC and the BNC. Therefore, we have similar ethical concerns as mentioned in that previous work, including the way the BNC data was collected. Those who so wish can easily exclude the BNC data (it comprises only a small part of the whole corpus) and retrain.

The RW-Eng corpus (v1 or v2) could introduce undesired bias in use outside the UK, since the data is sourced entirely from UK web pages and other UK sources from the 20th century. English used outside the UK, and more recent English anywhere, differ from this corpus in their word distributions, and therefore their input may yield sub-optimal or undesired results. Furthermore, models trained on it could encode a Western-centric view of the world.

The silver labels – the automatic parsing and tagging of the corpus – could introduce bias from the parsing/tagging algorithms. These parsers/taggers are also trained models, which could be affected by their data sources. If this is a concern for some users, we encourage them to perform validation of the data and its annotations.

Having said that, we believe that for most if not all conceivable applications, especially as long as one keep these limitations in mind, our work should not pose any practical risk.

References

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *ACL (1)*, pages 238–247.

Marco Baroni and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Klinton Bicknell, Jeffrey L Elman, Mary Hare, Ken McRae, and Marta Kutas. 2010. Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language*, 63(4):489–505.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, pages 69–72. Association for Computational Linguistics.

Grzegorz Chrupala. 2011. Efficient induction of probabilistic word classes with LDA. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 363–372, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Ronan Collobert and Jason Weston. 2007. Fast semantic extraction using a novel neural network architecture. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 25–32.

David Dowty. 1991. Thematic proto-roles and argument selection. *Language*, 67(3):547–619.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of english. *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

Todd R Ferretti, Ken McRae, and Andrea Hatherell. 2001. Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, 44(4):516–547.

Clayton Greenberg, Vera Demberg, and Asad Sayeed. 2015. Verb polysemy and frequency effects in thematic fit modeling. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 48–57, Denver, Colorado. Association for Computational Linguistics.

Luheng He, Kenton Lee, Omer Levy, and Luke Zettlemoyer. 2018. Jointly predicting predicates and arguments in neural semantic role labeling. *CoRR*, abs/1805.04787.

Xudong Hong, Asad Sayeed, and Vera Demberg. 2018. Learning distributed event representations with a multi-task approach. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 11–21, New Orleans, Louisiana. Association for Computational Linguistics.

Matthew Honnibal and Mark Johnson. 2015. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal. Association for Computational Linguistics.

Alessandro Lenci. 2011. Composing and updating verb argument expectations: A distributional semantic model. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, CMCL 2011, pages 58–66, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of distributional semantic models. *Language Resources and Evaluation*, pages 1–45.

Yuval Marton and Asad Sayeed. 2021. Thematic fit bits: Annotation quality and quantity for event participant representation.

Ken McRae, Michael J Spivey-Knowlton, and Michael K Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, 38(3):283–312.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*, volume 6, pages 2216–2219.

Ulrike Padó, Frank Keller, and Matthew W Crocker. 2006. Combining syntax and thematic fit in a probabilistic model of sentence processing. In *Proceedings of the 28th CogSci*, pages 657–662.

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

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.

Ottokar Tilk, Vera Demberg, Asad Sayeed, Dietrich Klakow, and Stefan Thater. 2016. Event participant modelling with neural networks. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 171–182, Austin, Texas. Association for Computational Linguistics.