Learning Distributed Event Representations with a Multi-Task Approach

Xudong Hong†, Asad Sayeed*, Vera Demberg†
†Dept. of Language Science and Technology, Saarland University
{xhong,vera}@coli.uni-saarland.de
*Dept. of Philosophy, Linguistics, and Theory of Science, University of Gothenburg
asad.sayeed@gu.se

Abstract

Human world knowledge contains information about prototypical events and their participants and locations. In this paper, we train the first models using multi-task learning that can both predict missing event participants and also perform semantic role classification based on semantic plausibility. Our best-performing model is an improvement over the previous state-of-the-art on thematic fit modelling tasks. The event embeddings learned by the model can additionally be used effectively in an event similarity task, also outperforming the state-of-the-art.

1 Introduction

Event representations consist, at minimum, of a predicate, the entities that participate in the event, and the thematic roles of those participants (Fillmore, 1968). *The cook cut the cake with the knife* expresses an event of cutting in which a cook is the "agent", the cake is the "patient", and the knife is the "instrument" of the action. Experiments have shown that event knowledge, in terms of the prototypical participants of events and their structured compositions, plays a crucial role in human sentence processing, especially from the perspective of *thematic fit*: the extent to which humans perceive given event participants as "fitting" given predicate-role combinations (Ferretti et al., 2001; McRae et al., 2005; Bicknell et al., 2010). Therefore, computational models of language processing should also consist of event representations that reflect thematic fit. To evaluate this aspect empirically, a popular approach in previous work has been to compare model output to human judgements (Sayeed et al., 2016).

The best-performing recent work has been the model of Tilk et al. (2016), who effectively simulate thematic fit via *selectional preferences*: generating a probability distribution over the full vocabulary of potential role-fillers. Given event context as input, including a predicate and a given set of semantic roles and their role-fillers as well as one target role, its training objective is to predict the correct role-filler for the target role. The objective of predicting upcoming role-fillers is cognitively plausible: there is ample evidence that humans anticipate upcoming input during sentence processing and learn from prediction error (Kuperberg and Jaeger, 2016; Friston, 2010) (even if other details of the implementation like back-propagation may not have much to do with how errors are propagated in humans).

An analysis of role filler predictions by Tilk et al.'s model shows that the model does not make sufficient use of the thematic role input. For instance, the representation of *apple eats boy* is similar to the representation of *boy eats apple*, even though the events are very dissimilar from one another. Interestingly, humans have been found to make similar errors. For instance, humans have been shown to frequently misinterpret a sentence with inverse role assignment, when the plausibility of the sentence with swapped role assignment is very high, as in *The mother gave the candle the daughter*, which is often erroneously interpreted as the daughter receiving the candle, instead of the literal syntax which says that the candle receives the daughter (Gibson et al., 2013).

Tilk et al.'s model design makes it more susceptible to this type of error than humans. The model lacks the ability to process in both directions, i.e., to both comprehend *and* produce thematic role marking (approximated here as thematic role assignment). We therefore propose to add a secondary role prediction task to the model, training it to both produce and comprehend language.

In this paper, we train the first model using multi-task learning (Caruana, 1998) which can ef-

*Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 11–21
New Orleans, June 5-6, 2018.

fectively predict semantic roles for event participants as well as perform role-filler prediction[1]. Furthermore, we obtain significant improvements and better-performing event embeddings by an adjustment to the architecture (parametric weighted average of role-filler embeddings) which helps to capture role-specific information for participants during the composition process. The new event embeddings exhibit state-of-the-art performance on a correlation task with human thematic fit judgements and an event similarity task.

Our model is the first joint model for selectional preferences (**SPs**) prediction and semantic role classification (**SRC**) to the best of our knowledge. Previous works used distributional similarity-based (Zapirain et al., 2013) or LDA-based (Wu and Palmer, 2015) SPs for semantic role labelling to leverage lexical sparsity. However, when it comes to a situation with domain shift, single task SP models that rely heavily on syntax have high generalisation error. We show that the multi-task architecture is better suited to generalise in that situation and can be potentially applied to improve current semantic role labelling systems which rely on small annotated corpora.

Our approach is a conceptual improvement on previous models because we address multiple event-representation tasks in a single model: thematic fit evaluation, role-filler prediction/generation, semantic role classification, event participant composition, and structured event similarity evaluation.

2 Role-Filler Prediction Model

Tilk et al. (2016) proposed a neural network, the non-incremental role-filler (**NNRF**) model, for role-filler prediction which takes a combination of words and roles as input to predict the filler of a target role. For example, the model would take "waiter/ARG0" and "serve/PRD" and target role "ARG1" as input and return high probabilities to words like "breakfast", "dish", and "drinks".

The original NNRF model can be seen in Figure 1 (excluding the part of the architecture shown in the red box). The input layer is a role-specific embedding tensor $\mathbf{T} \in \mathbb{R}^{|V|\times|R|\times d}$ that is indexed by two one-hot encoded vectors $\mathbf{w}_i$ and $\mathbf{r}_i$ for input word w_i and input role r_i, where V is the set of words and R is the set of semantic roles in our vocabulary. Tilk et al. applied *Tensor Factorisation*, which reduces the number of parameters to $(|V| + |R| + d) \times k$. The embedding tensor is factorised into three matrices[2]: $\mathbf{A}_e \in \mathbb{R}^{|V|\times k}$, $\mathbf{B}_e \in \mathbb{R}^{|R|\times k}$ and $\mathbf{C}_e \in \mathbb{R}^{k\times d}$. The overall embedding for a pair consisting of a word and its role, referred to as an *event participant embedding*, is represented as:

$$\mathbf{p}_l = (\mathbf{w}_i\mathbf{A}_e \circ \mathbf{r}_i\mathbf{B}_e)\mathbf{C}_e \quad (1)$$

where "$\circ$" is the Hadamard product.

When several word-role pairs $l = (w_i, r_i) \in C$, where C is the event context, are given as input, the model sums up their event participant embedding vectors to yield an *event representation* $\mathbf{e}$. Then it passes through one non-linearity layer with a parametric rectified linear unit (He et al., 2015): $\mathbf{h} = PReLU(\mathbf{e} + \mathbf{b}_e)$ where $\mathbf{b}_e$ is a bias vector.

The output layer consists of a softmax regression classifier computed as:

$$\mathbf{o}_w = Softmax_w(\mathbf{h}\mathbf{W}_w + \mathbf{b}_w) \quad (2)$$

where $\mathbf{b}_w$ is a bias vector. For each target role r_t, the model learns a target role-specific classifier with weight matrix of $\mathbf{W}_w^{(r_t)} \in \mathbb{R}^{d\times|V|}$, using r_t and event context C to predict the target word w_t. The weight matrices are stacked into an order-3 tensor and then factorised as:

$$\mathbf{W}_w^{(r_t)} = \mathbf{C}_w \; diag(\mathbf{r}_t\mathbf{B}_w) \; \mathbf{A}_w \quad (3)$$

where $diag(\mathbf{v})$ is a diagonal matrix with vector $\mathbf{v}$ on its main diagonal.

However, we found that the NNRF model in some cases relies heavily on lexical features but is not sensitive enough to semantic role assignments and hence represents phrases like "boy eats apple" in a similar way as "apple eats boy". We believe that a reason for this lies in the fact that the correct filler can often be predicted even when the role assignment is ignored, i.e., with the current objective, the model can often neglect the thematic role information. One could easily imagine that even humans might show similar behaviour if they only had to guess meanings from words they hear and are not required to produce correctly marked language themselves. We thus propose to add a second task to the network in order to approximate the dual comprehension and production tasks in human language learning.

[1]The source code and the supplemental document are available at https://github.com/tony-hong/event-embedding-multitask

[2]Further explanations are in the supplemental material.

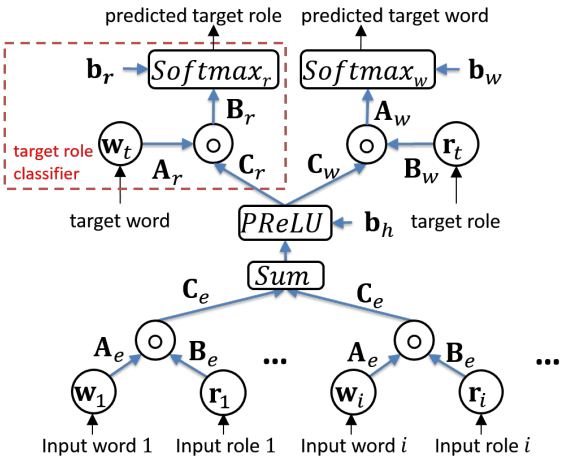

Figure 1: Architecture of multi-task role-filler model.

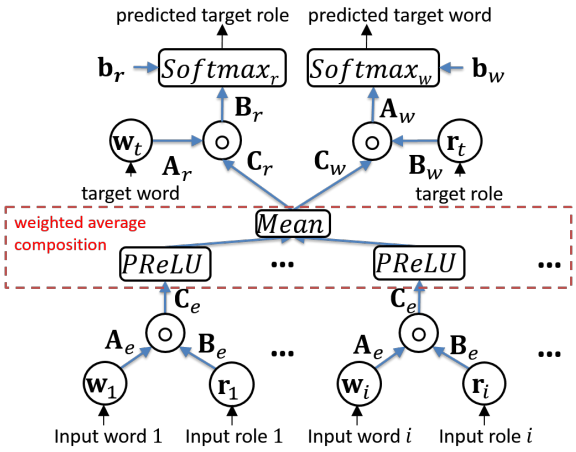

Figure 2: Architecture of role-filler averaging model.

3 Multi-Task Model

Our core idea is to add a second task, semantic role classification, such that the role-filler prediction model needs to predict the correct semantic role label for a target role-filler and a given set of input word-role pairs, i.e., the event context. Multi-task learning can integrate different objectives into one model and has previously been shown to help improve model generalisation (e.g., Caruana, 1998). The auxiliary task can be considered a regularisation of the objective function of the main task.

A neural model can be extended to multi-task architecture straightforwardly via sharing the low-level distributed representations. We design a multi-task model (**NNRF-MT**) which shares the event participant embedding for the event context and tackles role-filler prediction and semantic role classification simultaneously.

Figure 1 shows the NNRF-MT model with an additional role prediction classifier in the last layer, indicated by the red box. The new target role classifier mirrors the design of the original target word classifier. The output vector of the new target role classifier is computed as:

$$\mathbf{o}_r = Softmax_r(\mathbf{h}\mathbf{W}_r + \mathbf{b}_r) \quad (4)$$

where $\mathbf{W}_r \in \mathbb{R}^{d \times |R|}$ is the weight matrix of the target role classifier, and $\mathbf{b}_r$ is its bias vector. Like Equation (3), the weight matrix $\mathbf{W}_r^{(w_t)}$ for the target word w_t is factorised as:

$$\mathbf{W}_r^{(w_t)} = \mathbf{C}_r \, diag(\mathbf{w}_t \mathbf{A}_r) \, \mathbf{B}_r \quad (5)$$

where $\mathbf{A}_r \in \mathbb{R}^{|V| \times k^{(r)}}$.

3.1 Parametric Role-Filler Composition

In the NNRF-MT model, the embedding vectors of each word-role pair are simply summed up to represent the event. But in many cases, event participants contribute to the event differently. This has the disadvantage that some important participants are not correctly composed. Even worse, there is no normalization between cases where different numbers of role-filler pairs are available as context.

We thus propose a parametric architecture where PReLU is applied to each word-role pair embedding, and the resulting vectors are then combined by using the **mean** composition function. Parameters inside PReLU can now act as weights for each role-filler embedding. Computing the mean can be considered as the normalisation of role-filler representations within the event boundary, which can prevent the possible over-/underflow of the weights of the hidden vector.

With this method, the event embedding is computed as:

$$\mathbf{e} = \frac{1}{|C|} \sum_{l \in C} PReLU_l(\mathbf{p}_l) \quad (6)$$

and then directly fed into the classifier as the hidden vector $\mathbf{h}$. Figure 2 shows the resulting model named Role-Filler Averaging model (**RoFA-MT**), which is identical to the NNRF-MT model, except for the composition of event participant embeddings (marked by the red box).

3.2 Residual Learning

An additional way to reduce the challenge of exploding or vanishing gradients in factorised tensor is to apply *residual learning* (He et al., 2016).

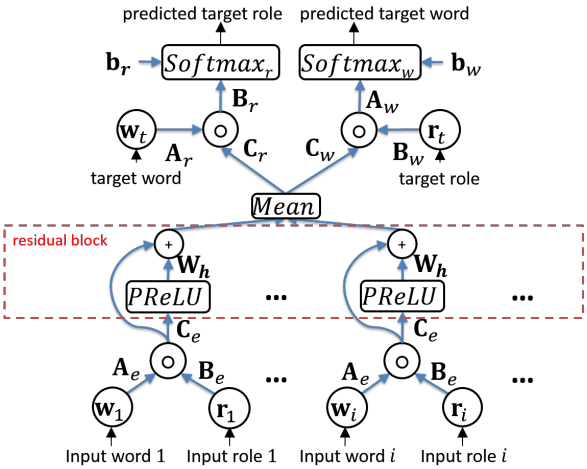

Figure 3: Architecture of residual role-filler averaging model.

The key idea in residual learning is that an identity mapping over other layers may be combined with a model that encodes information through several layers in order to simultaneously capture lower-level and higher-level information. We therefore experiment with residual learning in our RoFA-MT model (henceforth called **ResRoFA-MT**): the event participant vector now consists of a "raw" vector and a weighted vector that has been fed through a linear hidden layer, see Figure 3.

The original weight of the role-filler embedding is passed into the non-linear layer as: $\mathbf{h}_l = PReLU_l(\mathbf{r}_l\mathbf{C}_e)$ where $\mathbf{r}_l = \mathbf{w}_i\mathbf{A}_e \circ \mathbf{r}_j\mathbf{B}_e$ is the residual (i.e., the composition of word embedding and semantic role embedding). Then the combination of the output hidden vector $\mathbf{h}_l$ and the residual vector goes into the event embedding as:

$$\mathbf{e} = \frac{1}{|C|}\sum_{l \in C}(\mathbf{h}_l\mathbf{W}_h + \mathbf{r}_l) \quad (7)$$

where $\mathbf{W}_h$ is the weight matrix. After that, the event embedding goes directly into the classifier.

3.3 Multi-Task Training

The multi-task model is trained to optimise two objective functions in parallel. For each clause in the training data, we extract the predicate and all participants. We then choose each word-role pair as the target and the remainder as context C for one training sample. We use the multi-task model to predict the target role given the target filler as an input and to predict the target filler given the target role. We use a weighted combination of the probabilities of the target role and target word to obtain the overall loss function as:

$$\mathcal{L} = \mathcal{L}^{(w)}(C, r_t) + \alpha\mathcal{L}^{(r)}(C, w_t)$$

where α is the hyper-parameter of the weight of the semantic role classification task and can be tuned for different training purposes. In this paper, we choose 1.0 as the weight of semantic role prediction α to balance between two tasks.

4 Experiments

To learn an event representation from language resources with access to generalised event knowledge, we use the Rollenwechsel-English (**RW-eng**) corpus[3], a large-scale corpus based on BNC and ukWaC with about 2B tokens, which contains automatically generated PropBank-style semantic role labels for the head words of each argument (Sayeed et al., 2018).

We choose the first 99.2% as training data, the next 0.4% as validation data and the last 0.4% as test data, which follows Tilk's setting to make a fair comparison. From the training data, we extract a word list of the 50K most frequent head words (nouns, verbs, adjectives and adverbs) and add one OOV symbol[4]. For training the model, we distinguish between seven role labels: `PRD` for predicates, `ARG0`, `ARG1`, `ARGM-MNR`, `ARGM-LOC`, `ARGM-TMP`; all other roles are mapped onto a category `OTHER`.

NNRF is the current state-of-the-art model for event representation; we reimplement this model and use it as the baseline for our evaluation tasks. For a fair comparison, we train the NNRF model and our three multi-task models on the newest version of RW-eng corpus. Each model is trained for 27 iterations (or less if the model converged earlier)[5].

Because we use random parameter initialisation, to observe its effect to our evaluations, we train 10 instances of each model and report average performance (we do not use these 10 models as an ensemble method such as labelling by majority voting).

[3] `http://rollen.mmci.uni-saarland.de/RW-eng`

[4] A detailed description of data preprocessing is in the supplemental.

[5] The details of hyper-parameter setting are in the supplemental.

Model	Accuracy	p-value
NNRF-MT	89.1	-
RoFA-MT	94.8	< 0.0001
ResRoFA-MT	94.7	< 0.0001

Table 1: Semantic role classification results for the three multi-task architectures.

5 Evaluation: Semantic Role Classification

We begin by testing the new component of the model in terms of how well the model can predict semantic role labels.

5.1 Role Prediction Given Event Context

We evaluate our models on semantic role prediction accuracy given the predicate and other arguments with their roles on the test dataset of the RW-eng corpus. Table 1 shows that the RoFA-MT and ResRoFA-MT models outperform the NNRF-MT model by a statistically significant margin (tested with McNemar's test), showing that the parametric weighted average composition method leads to significant improvements.

5.2 Classification for Verb-Head Pairs

Semantic role classification systems make heavy use of syntactic features but can be further improved by integrating models of selectional preferences (Zapirain et al., 2009). Here we compare the semantics-based role assignments produced by our model to predictions made by various selectional preference (SP) models in the first evaluation of Zapirain et al. (2013). E.g., the model is to predict `ARG1` for the pair (eat_{verb}, *apple*) without any other feature.

Zapirain et al. (2013) combined a verb-role SP model built on training data and an additional distributional similarity model trained on a large scale corpus for estimating the fit between verbs and their arguments for different roles. These thematic fit estimates are used to select the best role label for each predicate-argument pair.

We consider only following best variants as baselines:

Zapirain13^{Pado07}**:** This variant uses a distributional similarity model constructed on a general corpus (BNC) with Padó and Lapata (2007)'s syntax-based method.

Zapirain13$^{Lin98}_{in-domain}$**:** This variant contains Lin (1998)'s distributional similarity model which uses syntax-based clustering. The model is precomputed on a mixed corpus (including WSJ) which is in the same domain as the WSJ test set.

We apply our trained role labelling model directly to the test set, without touching the WSJ training/validation set. Following the baselines, for semantic roles which are not represented in our model, we do not make any prediction (this is reflected in lower recall for those cases).

The model is evaluated on the data set from the CoNLL-2005 shared task (Carreras and Màrquez, 2005), which contains the WSJ corpus as part of its training/validation/test sets and the Brown corpus as an out-of-domain test set (marked in Table 2 as **Brown**). We estimate 99% confidence intervals using the bootstrapping method, with 100 replications. We also construct a trivial baseline model, the ZeroR classifier, which predicts the majority class `ARG1` given any input.

Table 2 shows that the baseline model using Lin's similarities (Zapirain13$^{Lin98}_{in-domain}$) works best on the WSJ test dataset, statistically significantly outperforming each of the other methods ($p < 0.01$). However, this can be explained by the fact that this model is using semantic similarities obtained from the same domain as the WSJ test set. Among the models without using in-domain semantic similarity, ResRoFA-MT is significantly better than all others ($p < 0.01$).

On the Brown data, which is out-of-domain for all models, the ResRoFA-MT model achieves the best result and outperforms previous baselines significantly ($p < 0.01$). Without any training on the WSJ corpus, our best model has a much smaller gap between test and ood dataset (only about 3 F_1 points), which indicates that our multi-task models generalise better than previous baselines.

5.3 End-to-End Semantic Role Labelling

Future work will need to investigate in more detail whether the multi-task models proposed here can be used to improve the performance of existing semantic role labellers. While our model cannot be directly applied to a standard semantic role labelling task (because it assigns roles only to head words), we were able to combine the model with an existing semantic role labeller and obtained promising results. Adding embeddings based on the predicate and target word $\mathbf{hC}_r diag(\mathbf{w}_t\mathbf{A}_r)$ from the NNRF-MT model (see Equation (4), (5)) as a feature to the MATE semantic role labeller

Model	In domain: WSJ test P	R	F_1	Out-of-domain: Brown P	R	F_1	F_1^{test} - F_1^{ood}
ZeroR baseline	36.11	36.11	36.11	32.46	32.46	32.46	3.65
Zapirain13^{Pado07}	53.13	50.44	51.75	43.24	35.27	38.85	12.90
Zapirain13$^{Lin98}_{in-domain}$	59.93	**59.38**	**59.65****	50.79	48.39	49.56	10.09
NNRF-MT	55.80	49.16	52.27	53.43	45.42	49.10	3.17
RoFA-MT	67.93	51.19	58.39	65.71	47.36	55.05	3.34
ResRoFA-MT	**68.03**	51.27	58.47	**66.39**	**47.85**	**55.62****	2.85

Table 2: Results of semantic role classification given verb-head pairs. P is precision, R is recall and F_1 is F-measure. F_1 values with a mark are significantly higher than all other values in the same column, where (**) $p < 0.01$.

(Björkelund et al., 2010; Roth and Woodsend, 2014) leads to a small but statistically significant improvement of 0.11 points in F_1 score on the out-of-domain dataset used in the CoNLL-2009 semantic role labelling task (Hajič et al., 2009).

6 Evaluation: Thematic Fit Modelling

Next, we evaluate our multi-task models against human thematic fit ratings in order to assess whether the inclusion of the multi-task architecture leads to improvements on this task, following Padó et al. (2009); Baroni and Lenci (2010); Greenberg et al. (2015b); Sayeed et al. (2016).

6.1 Datasets

The human judgement data consists of verbs, a verbal argument with its role, and an average fit judgement score on a scale from 1 (least common) to 7 (most common), e.g., $ask, police/\texttt{AGENT}, 6.5$. We used:

Pado07: the dataset proposed by Pado (2007) consists of 414 predicate-participant pairs with judgements. The roles are agent and patient.

McRae05: the dataset from McRae et al. (2005) contains 1444 judgements of verbs with an agent or patient.

Ferretti01: the dataset proposed by Ferretti et al. (2001) contains 274 ratings for predicate-location pairs (**F-Loc**) and 248 rating for predicate-instrument pairs (**F-Inst**).

GDS: the dataset from Greenberg et al. (2015a) contains 720 ratings for predicates and patients.

6.2 Baseline Models

We compare our models against previous distributional semantic models used for thematic fit tasks; many of these are from the Distributional Memory (DM) framework (Baroni and Lenci, 2010) whose tensor space is a high-dimensional count space of verb-noun-relation tuples from a large-scale mixed corpus smoothed by local mutual information. The key idea in applying DM models to the thematic fit rating task is to construct a "prototype filler", and compare candidate fillers against the prototype using cosine similarity. The baseline models we compare against include NNRF and:

TypeDM: This is best-performing DM model from Baroni and Lenci (2010). Relations of verb-noun pairs are obtained using hand-crafted rules. The results of this model are from reimplementations in Greenberg et al. (2015a,b).

SDDM-mo: This DM comes from Sayeed and Demberg (2014) and is constructed with automatically-extracted semantic information.

GSD15: This is the overall best-performing model from Greenberg et al. (2015b) using hierarchical clustering of typical role-fillers to construct prototype on TypeDM.

SCLB17: This is the best-performing model on F-Inst from Santus et al. (2017). The number of fillers used in prototype construction is 30 and the number of top features is 2000. We report the highest results among the different types of dependency contexts in their framework.

6.3 Methods and Results

We correlated the human judgements with the output probability of the role-filler given the predicate and the role. To avoid conflation between frequency in the training dataset and plausibility of the role-filler, we adopt the practice proposed in Tilk et al. (2016) to set the bias of the output layer to zero during the evaluation. We consider the NNRF model as our baseline and perform a two-tailed t-test to calculate statistical significance between the baseline model and each of the three models proposed in this paper.

Model	Pado07	McRae05	F-Loc	F-Inst	GDS	avg
TypeDM	53	33	23	36	46	40.8
SDDM-mo	**56**	27	13	28	-	-
GSD15	50	36	29	42	48	40.5
SCLB17	49	28	37	50	-	-
NNRF	43.3	35.9	**46.5**	**52.1**	57.6	44.2
NNRF-MT	43.2	36.1	46.3	50.0*	57.2	44.0
RoFA-MT	52.2**	41.9**	45.9	49.4*	60.7**	48.6**
ResRoFA-MT	53.0**	**42.5****	46.3	47.7**	**60.8****	**48.9****

Table 3: Results on human thematic fit judgement correlation task (Spearman's $\rho \times 100$) compared to previous work. The last column reports the weighted average results by numbers of entries of all five datasets. Values with a mark are significantly different from the baseline model (NNRF), where (*) $p < 0.05$, (**) $p < 0.01$.

Table 3 shows results for all models and datasets. The ResRoFA-MT model performs best overall, improving more than 4 points over the baseline. The multi-task model (NNRF-MT) has performance similar to baseline (NNRF). Our new architecture using a parametric weighted average over event participant embeddings (RoFA-MT) outperforms simple summation (NNRF-MT), especially on the Pado07, McRae05 and GDS datasets. The residual method leads to further minor improvements on the Pado07, F-Loc and GDS datasets. However, on predicate-instrument pairs of the F-Inst dataset, NNRF outperforms other models significantly. We think that multi-task models are biased towards roles with a larger frequency like `ARG0` or `ARG1`, which is proved in the ablation study (see Section 8).

7 Evaluation: Compositionality

The thematic fit judgements from the tasks discussed in section 6 only contain ratings of the fit between the predicate and one role-filler. However, other event participants contained in a clause can affect human expectations of the upcoming role-fillers. For instance, mechanics are likely to check tires, while journalists are likely to check spellings. The **B10** dataset (Bicknell et al., 2010) contains human judgements for 64 pairs of agent-verb-patient triples, where one triple in each pair is plausible (e.g., "journalist check spelling"), and one is implausible (e.g., "journalist check type"). A model is evaluated based on whether it successfully assigns a higher likelihood/rating to the plausible than to the implausible object (also referred to as the *Accuracy 1* metric in Tilk et al. (2016)).

The baseline models are NNRF as well as:

Random: The naive baseline model consists of choosing the tags uniformly at random.

Lenci11: Lenci (2011) proposed a composition model for TypeDM.

Table 4 shows that our new composition method based on parametric weighted average outperforms previous models; the RoFA-MT model achieves the highest accuracy overall and outperforms the baseline (NNRF) significantly.

7.1 Event Similarity

Lastly, we evaluate the quality of the event embeddings learned via the multi-task network models. While word embeddings from tools like word2vec (Mikolov et al., 2013) are standard methods for obtaining word similarities, identifying a suitable method for more general event similarity estimation is still a relevant problem. The model proposed here constitutes an interesting method for obtaining event embeddings, as it is trained on two semantics-focused prediction tasks.

For evaluation, we use the sentence similarity task proposed by Grefenstette and Sadrzadeh (2015) (second experiment in their paper). For evaluation, we use the re-annotated dataset, named **GS13**, constructed in 2013 by Kartsaklis and Sadrzadeh (2014). Each row in the dataset contains a participant ID, two sentences, a human evaluation score of their similarity from 1 to 7, and a `HIGH`/`LOW` tag indicating the similarity group of two sentences. An example entry is:

p1, (table, draw, eye), (table, attract, eye), 7, `HIGH`

where p1 is the participant ID. We compare our models' performance to NNRF, as well as:

Kronecker: The best-performing model in Grefenstette and Sadrzadeh (2015) using Kronecker product as its composition method.

W2V: The sentence representations in W2V are

	Random	Lenci11	NNRF	NNRF-MT	RoFA-MT	ResRoFA-MT
Accuracy 1	0.50	0.67	0.73	0.71	**0.76***	0.75

Table 4: Results on agent-patient compositionality evaluation comparing to previous models. Values with a mark are significantly different from the baseline model (NNRF), where (*) $p < 0.05$.

	W2V	Kronecker	NNRF	NNRF-MT	RoFA-MT	ResRoFA-MT	Human
$\rho \times 100$	13	26	34.2	35.7	34.0	**36.7****	60

Table 5: Results on event similarity evaluation comparing to previous models. Values with a mark are significantly different from the baseline model (NNRF), where (*) $p < 0.05$, (**) $p < 0.01$.

constructed by element-wise addition of pre-trained word2vec (Mikolov et al., 2013) word embeddings.

Human: Mean inter-annotator correlation using Spearman's ρ. This can be considered to be the upper bound of the task.

To estimate sentence similarity, we feed all three words and their roles (`ARG0`/`PRD`/`ARG1`) into each model. We then extract the event representation vectors for both sentences and compute their cosine similarity. Table 5 shows correlation coefficients in Spearman's $\rho \times 100$ between sentence-pair similarities and human judgement scores. ResRoFA-MT obtains best results, indicating that the secondary task helped also to improve the network-internal event representations. These results indicate that ResRoFA-MT-based event embeddings may be suitable for applications and tasks where similarity estimates for larger phrases are needed (cf. Wanzare et al., 2017).

8 Ablation Study: Single-task Variants

From the evaluations above, we notice that the performance of the multi-task model with simple addition composition method (NNRF-MT) is not significantly different from the single task model (NNRF). In order to test whether the additional training task improves model performance, we develop single-task variants for RoFA-MT and ResRoFA-MT models, named RoFA-ST and ResRoFA-ST correspondingly, by taking out the semantic role classifiers. We then perform one-trial experiments and evaluate the models on thematic fit modelling and compositionality tasks by comparing the one-trial results of single-task variants versus the confidence intervals obtained from the 10 runs of the multi-task models.

The results in Table 6 show that multi-task models significantly outperform single-task models on Pado07, McRae05, F-Loc, GDS and overall. However, single-task variants are superior to multi-task models on F-Inst dataset, which is consistent with our findings in Section 6.2. On the compositionality tasks, the multi-task architecture improves only the performance of the residual weighted average model (ResRoFA-MT) but harms the event similarity performance of the weighted average model (RoFA-MT).

9 Related Work

Modi et al. (2017) proposed a compositional neural model for referent prediction in which event embeddings were constructed via the sum of predicate and argument embeddings. Weber et al. (2017) proposed a tensor-based composition model to construct event embeddings with agents and patients. They represented predicates as tensors and arguments as vectors. Cheng and Erk (2018) proposed a neural-based model to predict implicit arguments with event knowledge in which the event embeddings are composed with a two-layer feed-forward neural network.

10 Conclusions

This paper introduced two innovations to the modelling of events and their participants at the clause level: (1) we proposed a multi-task model of role-filler prediction and semantic role classification; (2) we proposed a parametric weighted average method which improves the composition of event participants in the input.

The introduction of semantic role classification as a secondary task addressed a weakness of Tilk et al. (2016)'s model. The semantic role classification task requires a much stronger internal representation of the semantic roles on top of lexical information. Thanks to the internal hidden layer shared between the two tasks, the event representation profited from the additional learning objective, increasing the models' performance on esti-

Model	Pado07	McRae05	F-Loc	F-Inst	GDS	avg	B10	GS13
RoFA-ST	44.1***	36.6***	44.4*	**56.7*****	57.3***	44.5***	75.0	36.3**
RoFA-MT	52.2	41.9	45.9	49.4	60.7	48.6	**76.1**	34.0
ResRoFA-ST	42.3***	35.8***	44.5**	50.4*	56.9***	43.6***	67.2***	32.5***
ResRoFA-MT	**53.0**	**42.5**	**46.3**	47.7	**60.8**	**48.9**	74.5	**36.7**

Table 6: Ablation study of single task variants. Underlined values indicate the best values with the same composition method, and bold values indicate the best values on that data set. Values with a mark are significantly different from the multi-task baseline models (RoFA-MT / ResRoFA-MT), where (*) $p < 0.05$, (**) $p < 0.01$, (***) $p < 0.001$.

mating event similarity.

We also performed a study regarding the usefulness of our purely semantics-based representations for semantic role labelling. While many semantic role labellers rely predominantly on syntax, our approach addresses the likelihood that a semantic role should be assigned purely based on its plausibility to fill that role content-wise. We showed that the semantics-based role label predictions generated by our multi-task model outperform the ones based on earlier syntax-based selectional preference methods and observe promising results for integrating the model with a semantic role labeller on out-of-domain data.

Our parametric composition method (RoFA-MT) composes event embeddings in the hidden layer, which captures role-specific information during the composition process and reduces the risk of overflow and underflow of the hidden layer weights. We additionally included the residual learning method alongside RoFA-MT (ResRoFA-MT), further mitigating the vanishing/exploding gradient problem and allowing the transmission of information from lower levels directly into the event embedding. This approach provided the overall best result of all models on the thematic fit human judgement task as well as the event similarity task and competitive results on the tasks individually.

10.1 Future Work

In future work, the model may be improved by including visual information from photos and videos. Common-sense reasoning is becoming a new focus (Mostafazadeh et al., 2016; Baroni et al., 2017). One characteristic of common-sense knowledge is that it is often not explicitly mentioned in language precisely *because* it constitutes common-sense knowledge and is hence uninformative as it can easily be inferred (Mukuze et al., 2018). Syntactically optional event participants (such as the kitchen as location for the predicate "cook") are thus often omitted in text; this sets a limit to what can be learned from text only.

The prospect of applying our models independently to SRL tasks suggests an area of potential future work. Our models currently use only the predicates and head words of arguments. Instead of depending on corpora with extracted head words, we can integrate an attention mechanism (Vaswani et al., 2017) to capture the position of syntactic heads. We are working on extending our models to use all words, which will enable testing as an SRL tool.

Finally, the predictive nature of this type of model can potentially enable its deployment in incremental semantic parsing (Konstas et al., 2014; Konstas and Keller, 2015) by combining the multi-task design with the incremental architecture in (Tilk et al., 2016). We are continuing to develop this and other ways of employing models of event representation that simultaneously predict event participants and assess the fit of given participants.

Acknowledgements

This research was funded in part by the German Research Foundation (DFG) as part of SFB 1102 "Information Density and Linguistic Encoding", EXC 284 Cluster of Excellence "Multimodal Computing and Interaction", and a Swedish Research Council (VR) grant for the Centre for Linguistic Theory and Studies in Probability (CLASP). We thank Dr. Michael Roth for conducting the semantic role labelling evaluation using features from our model to see whether it is beneficial. We sincerely thank the anonymous reviewers for their insightful comments that helped us to improve this paper.

References

Marco Baroni, Armand Joulin, Allan Jabri, Germàn Kruszewski, Angeliki Lazaridou, Klemen Simonic, and Tomas Mikolov. 2017. CommAI: Evaluating the first steps towards a useful general AI. *arXiv preprint arXiv:1701.08954* .

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.

Anders Björkelund, Bernd Bohnet, Love Hafdell, and Pierre Nugues. 2010. A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*. Association for Computational Linguistics, pages 33–36.

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

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

Pengxiang Cheng and Katrin Erk. 2018. Implicit argument prediction with event knowledge. *arXiv preprint arXiv:1802.07226* .

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.

C. J. Fillmore. 1968. The case for case. *Universals in Linguistic Theory* pages 1–25(Part Two).

Karl Friston. 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11(2):127.

Edward Gibson, Leon Bergen, and Steven T Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences* 110(20):8051–8056.

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

Clayton Greenberg, Asad Sayeed, and Vera Demberg. 2015b. Improving unsupervised vector-space thematic fit evaluation via role-filler prototype clustering. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 21–31.

Edward Grefenstette and Mehrnoosh Sadrzadeh. 2015. Concrete models and empirical evaluations for the categorical compositional distributional model of meaning. *Computational Linguistics* 41(1):71–118.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*. Association for Computational Linguistics, Boulder, Colorado, pages 1–18.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. pages 1026–1034.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 770–778.

Dimitri Kartsaklis and Mehrnoosh Sadrzadeh. 2014. A study of entanglement in a categorical framework of natural language. *arXiv preprint arXiv:1405.2874* .

Ioannis Konstas and Frank Keller. 2015. Semantic role labeling improves incremental parsing. 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)*. Association for Computational Linguistics, Beijing, China, pages 1191–1201.

Ioannis Konstas, Frank Keller, Vera Demberg, and Mirella Lapata. 2014. Incremental semantic role labeling with Tree Adjoining Grammar. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 301–312.

Gina R Kuperberg and T Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Language, cognition and neuroscience* 31(1):32–59.

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. Association for Computational Linguistics, Stroudsburg, PA, USA, CMCL 2011, pages 58–66.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*. Association for Computational Linguistics, pages 768–774.

Ken McRae, Mary Hare, Jeffrey L Elman, and Todd Ferretti. 2005. A basis for generating expectancies for verbs from nouns. *Memory & Cognition* 33(7):1174–1184.

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*. pages 3111–3119.

Ashutosh Modi, Ivan Titov, Vera Demberg, Asad Sayeed, and Manfred Pinkal. 2017. Modelling semantic expectation: Using script knowledge for referent prediction. *Transactions of the Association of Computational Linguistics* 5(1):31–44.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696* .

Nelson Mukuze, Anna Rohrbach, Vera Demberg, and Bernt Schiele. 2018. A vision-grounded dataset for predicting typical locations for verbs. In *The 11th edition of the Language Resources and Evaluation Conference (LREC)*.

Sebastian Padó and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics* 33(2):161–199.

Ulrike Pado. 2007. *The Integration of Syntax and Semantic Plausibility in a Wide-Coverage Model of Human Sentence Processing*. Ph.D. thesis, Saarland University.

Ulrike Padó, Matthew W Crocker, and Frank Keller. 2009. A probabilistic model of semantic plausibility in sentence processing. *Cognitive Science* 33(5):794–838.

Michael Roth and Kristian Woodsend. 2014. Composition of word representations improves semantic role labelling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pages 407–413.

Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Measuring thematic fit with distributional feature overlap. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 659–669.

Asad Sayeed and Vera Demberg. 2014. Combining unsupervised syntactic and semantic models of thematic fit. In *Proceedings of the first Italian Conference on Computational Linguistics (CLiC-it 2014)*.

Asad Sayeed, Clayton Greenberg, and Vera Demberg. 2016. Thematic fit evaluation: an aspect of selectional preferences. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*. Association for Computational Linguistics, Berlin, Germany, pages 99–105.

Asad Sayeed, Pavel Shkadzko, and Vera Demberg. 2018. Rollenwechsel-English: a large-scale semantic role corpus. In *Proceedings of the 11th edition of the Language Resources and Evaluation Conference*.

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*. Association for Computational Linguistics, Austin, Texas, pages 171–182.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. pages 6000–6010.

Lilian DA Wanzare, Alessandra Zarcone, Stefan Thater, and Manfred Pinkal. 2017. Inducing script structure from crowdsourced event descriptions via semi-supervised clustering. *LSDSem 2017* page 1.

Noah Weber, Niranjan Balasubramanian, and Nathanael Chambers. 2017. Event representations with tensor-based compositions. *arXiv preprint arXiv:1711.07611* .

Shumin Wu and Martha Palmer. 2015. Can selectional preferences help automatic semantic role labeling? In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*. pages 222–227.

Beñat Zapirain, Eneko Agirre, and Lluís Màrquez. 2009. Generalizing over lexical features: Selectional preferences for semantic role classification. In *Proceedings of the ACL-IJCNLP 2009 conference short papers*. Association for Computational Linguistics, pages 73–76.

Benat Zapirain, Eneko Agirre, Lluis Marquez, and Mihai Surdeanu. 2013. Selectional preferences for semantic role classification. *Computational Linguistics* 39(3):631–663.