Joint Modeling of Arguments for Event Understanding

Yunmo Chen **Tongfei Chen** **Benjamin Van Durme**
Johns Hopkins University
{yunmo,tongfei,vandurme}@jhu.edu

Abstract

We recognize the task of event *argument linking* in documents as similar to that of intent *slot resolution* in dialogue, providing a Transformer-based model that extends from a recently proposed solution to resolve references to slots. The approach allows for joint consideration of argument candidates given a detected event, which we illustrate leads to state-of-the-art performance in multi-sentence argument linking. [1]

1 Introduction

Given an event recognized in text, we are concerned with finding its associated arguments. Significant work has focused at the level of single sentence contexts, such as in *semantic role labeling* (SRL; Gildea and Jurafsky, 2000; He et al., 2017; Ouchi et al., 2018, *inter alia*). Unfortunately even perfect performance in SRL will be limited by the existence of arguments outside the sentence boundary, leading to prior work (Das et al., 2010; Silberer and Frank, 2012; Ebner et al., 2020) on an alternative paradigm variously called *implicit role resolution* or *argument linking*, where an event trigger (e.g. "attack") evokes a set of roles (e.g. ATTACKER, TARGET) to be filled, and they are linked to explicit argument mentions found in text. In argument linking, possible candidate arguments are first detected, then linked to specific roles of detected events. This bears similarity to coreference resolution, where document-level context can be aptly utilized. For an example, see Figure 1.

This formulation is similar to the resolution of referring expressions in conversational dialogues (Çelikyilmaz et al., 2014), where a current utterance is considered to invoke an *intent* (e.g. BUY-BOOK), accompanied by a number of *slots* (e.g. NAME, AUTHOR, PUBLISHER, etc.). Even more than in event argument linking, in dialogue systems the sentence-level (utterance-level) context often fails to contain all salient arguments (slots): slots from previous rounds of dialogue may often be relevant to the current intent.[2]

Dialogue	Events
Intent type BUY-BOOK	Event type ATTACK
Slot key NAME, AUTHOR	Role type ATTACKER, TARGET
Slot value *1984*, *George Orwell*	Argument *Russia*, *Ukraine*

Table 1: Mapping between terminologies in intent slot resolution and event argument linking, with examples.

We propose a novel model for joint modeling of potential arguments inspired by Chen et al. (2019) for slot-filling in dialogue systems, which proposed to jointly predict spans that are relevant to the intent of the current round of dialogue. Over detected arguments, a Transformer (Vaswani et al., 2017) encoder is placed upon the event trigger and potential *arguments* to jointly learn the relations between the event trigger and its arguments. The input to this Transformer is no longer *tokens* but *spans*: given the Transformer output of each span, a classification loss is utilized to perform argument role classification. We demonstrate this leads to state-of-the-art performance on the RAMS argument linking dataset introduced by Ebner et al. (2020),[3] showing the benefits of joint modeling when linking arguments to roles of events.

[1] Our code can be found at https://github.com/wanmok/joint-arglinking.

[2] E.g., from Chen et al. (2019): *What's the weather in San Francisco? ... Any good Mexican restaurants there?*

[3] https://nlp.jhu.edu/rams.

Proceedings of the First Workshop on Computational Approaches to Discourse, pages 96–101
Online, November 20, 2020.
https://doi.org/10.18653/v1/P17

2 Background

Implicit role resolution Palmer et al. (1986) treated unfilled semantic roles as special cases of anaphora and coreference resolution. Starting from the SemEval 2010 Task 10: Linking Roles (Ruppenhofer et al., 2010), there have been more recent modeling efforts on this task. Chen et al. (2010) approached this with their SRL system SEMAFOR (Das et al., 2010), casting the task as extended SRL by admitting constituents (potential arguments) from context larger than sentence boundaries. Silberer and Frank (2012) considered the problem as an anaphora resolution task within the discourse context. Ebner et al. (2020) similarly considered the task as related to anaphora resolution, and introduced a new dataset, RAMS, for exploring non-local argument linking. See O'Gorman (2019) and Ebner et al. (2020) for further background.

Event extraction In event extraction there are historically three subtasks: detecting event triggers, detecting entity mentions, and then *argument role prediction*, where relations between mentions and triggers are predicted in accordance to the event type's predefined set of roles under a closed ontology. Prior work has proposed pipeline system of the subtasks (Ji and Grishman, 2008; Li et al., 2013; Yang and Mitchell, 2016, *inter alia*), or as a joint model over the three tasks (Nguyen and Nguyen, 2019; Lin et al., 2020, *inter alia*). Our work could be seen as a version of argument role prediction, but which operates beyond sentence boundaries.

Frame-based SLU In dialogue systems, semantic frame based spoken language understanding (SLU) is one of the most commonly applied SLU technologies for human-computer interaction. Such systems often output an interpretation of dialogues represented as *intents* and *slots* (Wang et al., 2011). Çelikyilmaz et al. (2014) and Bapna et al. (2017) proposed models to resolve references to slots in the dialogue, tracking conversation states across multiple dialogue turns. Dhingra et al. (2017) augmented such methods with external knowledge bases (KBs) to create a multi-turn dialogue agent which helps users search KBs. Chen et al. (2019) proposed joint models over potential slots in dialogue to output which contextual slots should be carried over to the most recent utterance. Our approach is inspired by this work, by drawing analogies between concepts in SLU (intents / slots) and those in IE (events / arguments) (see Table 1).

3 Problem Formulation

Following Ebner et al. (2020) we consider argument linking as the task of choosing amongst detected mention span candidates given detected event trigger spans. Given a document $d = (w_1, \cdots, w_n)$ where each w_i is a word, entity mention set M (candidate arguments) containing mentions $m_i = d[l_i : r_i] \in M$ where l_i and r_i demarcates the left and right boundary (both inclusive), and a event trigger span $t = d[l_t : r_t]$, an argument linking model predicts the role (or absence) of each mention with respect to the event.

An event ontology can be formulated as a set of event types $\mathcal{T}$, where each type $e \in \mathcal{T}$ is associated with a set of *roles* $R(e)$,[4] while other roles are non-permissible. We denote the union of all roles for all event types, plus an empty ε role (a dummy role denoting an argument is not part of the event structure) as $\mathcal{R} = \bigcup_{e \in \mathcal{T}} R(e) \cup \{\varepsilon\}$.

4 Approach

Argument and trigger representation We compute a fixed-length vector with dimension d for each argument and trigger span as their representations. To compute this, we first pass the document through a pre-trained contextualizing model (BERT (Devlin et al., 2019) here).[5] We split documents into sentences and feed each sentence to BERT for encoding. Each token w_i might be split into more than 1 subword units—in this case we take the average of these subword representations so that each token w_i has 1 vector representation $\mathbf{w}_i \in \mathbb{R}^{d_{\text{tok}}}$, following Zhang et al. (2019).

For an argument span $m = (w_l, \cdots, w_r)$, we follow Lee et al. (2017) to generate a span embedding.[6] The span embedding $\mathbf{m}$ for mention span m comprises of three parts, the representation of its left boundary, its right boundary, and a learned pooling over the tokens in the span. This learned pooling utilized a global attention query vector $\mathbf{q} \in \mathbb{R}^{d_{\text{tok}}}$, and computes the weighted sum of all tokens with respect to the attention scores derived from $\mathbf{q}$:

$$a_i = \frac{\exp \mathbf{q}^{\mathrm{T}} \mathbf{w}_i}{\sum_{j=l}^{r} \exp \mathbf{q}^{\mathrm{T}} \mathbf{w}_j}; \quad \mathbf{c} = \sum_{i=l}^{r} a_i \cdot \mathbf{w}_i, \quad (1)$$

[4] For example, in the ACE 2005 dataset, $R(\text{ATTACK}) = \{\text{ATTACKER}, \text{TARGET}, \text{INSTRUMENT}, \text{TIME}, \text{PLACE}\}$.

[5] Documents are chuncked into max-length 512 segments while respecting sentence boundaries, and each is fed to BERT respectively.

[6] The width embeddings in Lee et al. (2017) are not used.

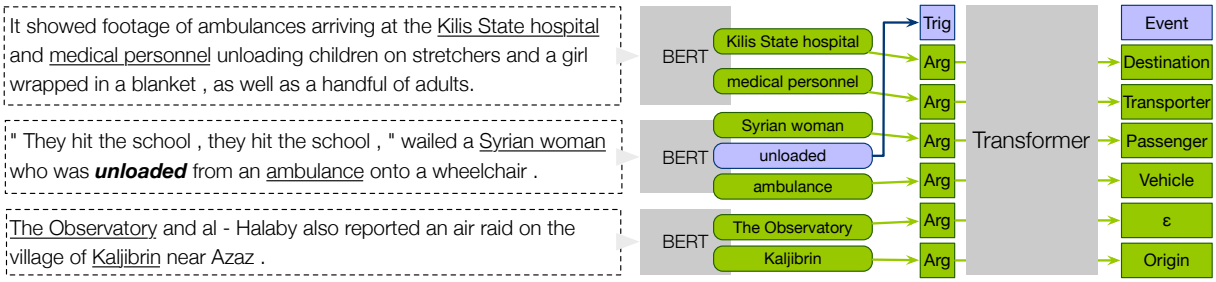

Figure 1: An example of our model running over a paragraph. Trigger and argument span representations are computed from BERT, then later fed to a Transformer for jointly modeling the spans to predict their roles.

and pass that through a 2-layer feed-forward neural network to yield a fixed-length vector $\mathbf{m}_i \in \mathbb{R}^{d_{\text{span}}}$ for each argument span m_i:

$$\mathbf{m} = \text{FFNN}_{\text{arg}}\left([\mathbf{w}_l\ ;\ \mathbf{w}_r\ ;\ \mathbf{c}]\right)\ . \tag{2}$$

Similarly, for any trigger span $t = [l : r]$, we employ a different set of parameters:

$$\mathbf{t} = \text{FFNN}_{\text{trig}}\left([\mathbf{w}_l\ ;\ \mathbf{w}_r\ ;\ \mathbf{c}]\right)\ . \tag{3}$$

Joint modeling of arguments We propose a joint model for all the arguments with respect to the given event trigger with event type e (see Figure 1). We form a sequence $(\mathbf{t}, \mathbf{m}_1, \mathbf{m}_2, \cdots, \mathbf{m}_n)$ with the trigger span encoding as the prefix, then followed by the representations of all the candidate mentions, then fed to a *Transformer* encoder (Vaswani et al., 2017). A Transformer, by its self-attention mechanism, naturally models the relation between every trigger-argument and argument-argument pair. Note two major differences as compared to a Transformer that runs on tokens: (1) each input to the Transformer represents a *span* instead of a token, following Chen et al. (2019); (2) since the arguments do not take an explicit sequential order, we forgo the positional embeddings in Transformers, effectively modeling the input as a *set* of spans instead of a *sequence* (self-attention exhibits the property of permutation invariance without positional embeddings (Lee et al., 2019)).

For each argument span input $\mathbf{m}_i$, we pass the output from the Transformer encoder $\hat{\mathbf{m}}_i$ to linear layer with the output size being the size of the role set $\mathcal{R}$. Softmax is applied to the output of size $|\mathcal{R}|$, with the non-permissible roles masked out, yielding a distribution over the set of roles designated by the given event type, plus the non-argument ε role:

$$P(r|t,m) = \frac{\exp \mathbf{w}_r^{\mathrm{T}} \hat{\mathbf{m}}}{\sum_{r' \in R(e) \cup \{\varepsilon\}} \exp \mathbf{w}_{r'}^{\mathrm{T}} \hat{\mathbf{m}}} \tag{4}$$

The model could hence be trained using a cross-entropy loss function to maximize such likelihood.

5 Experiments

As we draw the connections between SLU in dialogue systems and argument linking in information extraction, we focus primarily on evaluating the model a discourse-level dataset, RAMS (Ebner et al., 2020). First however we look at a more established dataset, ACE 2005 (Walker et al., 2006)[7], to verify if our model can reasonable performance compared to prior work in event understanding. While ACE 2005 is annotated only at the sentence-level, our model may still be applied in this setting. For detailed experimental setup, see Appendix A.

Baseline Aside from joint modeling of arguments, we also include an **independent** model as a case in ablation studies (while our proposed method labeled as **joint**). The independent model removes the Transformer encoder (cf. Equation 4), but directly applies a feed-forward neural network atop of the trigger representation and each argument representation to classify the role (or absence) of the argument with respect to the event trigger. [8]

$$P(r|t,m) = \frac{\exp \mathbf{w}_r^{\mathrm{T}} F_{\text{ind}}([\mathbf{t};\mathbf{m}])}{\sum_{r' \in R(e) \cup \{\varepsilon\}} \exp \mathbf{w}_{r'}^{\mathrm{T}} F_{\text{ind}}([\mathbf{t};\mathbf{m}])}$$

The result from model would show the difference between the proposed joint argument modeling approach v.s. a simpler, independent model.

[7] https://catalog.ldc.upenn.edu/LDC2006T06.

[8] This scoring function for triples (r, t, m) is similar to Ebner et al. (2020)'s model. However, their model is trained to maximize the posterior probability of the correct argument given a trigger and a role, whereas in our independent baseline here the probability of the correct role given a trigger and an argument candidate is maximized.

	Split	ACE 2005	RAMS
#Event types		33	139
#Role types		22	65
#Events/#Args	train	4202/4859	7329/17026
	dev	450/605	924/2188
	test	403/576	871/2023

Table 2: Dataset statistics.

Model	P	R	F_1
Lin et al. (2020)	*48.8*	*53.9*	*56.8**
Lin et al. (2020) PoE	-	-	*58.6**
Independent	48.0	76.7	59.0
Joint	56.0	79.2	65.6

Table 3: We verify our model achieves similar performance to recent work on ACE 2005. PoE denotes "product of experts", an ensemble model in Lin et al. (2020). * Results not directly comparable as we are exploring argument linking only.

Metrics We use precision, recall, and F_1-score as metrics. A link between the trigger and an argument is considered correct, if and only if the predicted argument span offsets and role matches the gold reference. We report using micro-average among F_1-scores across different roles.

5.1 ACE 2005

We use ACE 2005 as a sanity check for our discourse-context model to verify its ability to perform sentence-context extraction. We follow Lin et al. (2020)'s pre-processing and dataset splits for event extraction task (statistics see Table 2). Table 3 reports the experimental results on ACE 2005. Although the results are not directly comparable since our model has access to gold trigger/argument spans (Lin et al. (2020) does not), we can observe similar levels of performance, suggesting our method may be competitive when applied to event understanding beyond sentence boundaries.

5.2 RAMS

Roles Across Multiple Sentences (RAMS; Ebner et al., 2020) is an event extraction dataset that considers discourse-level, non-local arguments in document-level context. We follow the train/dev/test split provided in the dataset, with statistics shown in Table 2. Experiments setup follow the configuration employed for ACE 2005.

Table 4 shows the performance of our models on RAMS. Following the same conditions as Ebner et al. (2020), our joint model outperforms that work, and our independent baseline, by a substantial margin of 6.6%, illustrating the benefit of modeling potential arguments jointly.

Model	P	R	F_1
Ebner et al. (2020)	62.8	74.9	68.3
Ebner et al. (2020) TCD	78.1	69.2	73.3
Independent	73.5	73.0	73.3
Joint	**79.6**	**80.2**	**79.9**

Table 4: Experimental results on RAMS. TCD designates the use of ontology-aware type-constrained decoding, which is similar to our independent model.

Dist.	# Gold args.	RAMS-TCD	Ours
−2	79	75.7	**77.2**
−1	164	73.7	**74.4**
0	1,811	75.0	**79.6**
+1	87	76.5	**77.0**
+2	47	**79.1**	78.7

Table 5: Breakdown of the models' performance across sentence distances on the RAMS dev set. RAMS-TCD refers to Ebner et al. (2020)'s type-constrained decoding approach (see Table 4).

We analyze the performance of our model on non-local arguments, i.e., arguments that are not in the same sentence as the event trigger (Table 5). Our model's performance on non-local arguments is on par with local arguments, demonstrating the ability to handle non-local argument linking.

Case study We here show one example where the joint model performs better than the independent model. The joint model correctly labeled all the roles, while the independent model failed on two. We hypothesize that joint modeling of the arguments will avoid these cases where multiple spans are labeled with the same role.

... Stratfor analyst Sim Tack:" This was indeed an *Islamic State* **attack**, rather than an accidental *explosion*." New satellite imagery appears to reveal extensive damage to a strategically significant *airbase* in *central Syria* used by Russian forces ...

Argument	Independent	Joint	Gold
Islamic State	Attacker	Attacker	Attacker
explosion	~~Attacker~~	Instrument	Instrument
airbase	~~Attacker~~	Victim	Victim
central Syria	Place	Place	Place

6 Conclusion

We proposed a joint modeling approach for argument linking that considers the interdependent relationships among argument mentions conditioning on a specific event. Our approach extends from recent work in dialogue systems, viewing a document as essentially a single-side discourse, and where event arguments are recognized as similar to slots that potentially carryover across utterances. Experimental results show our approach achieves superior performance on a recently introduced dataset for modeling discourse-level contexts.

Acknowledgments

This research was supported by the JHU HLTCOE, DARPA AIDA, and IARPA BETTER. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes. The views and conclusions contained in this publication are those of the authors and should not be interpreted as representing official policies or endorsements of DARPA or the U.S. Government.

References

Ankur Bapna, Gökhan Tür, Dilek Hakkani-Tür, and Larry P. Heck. 2017. Sequential dialogue context modeling for spoken language understanding. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 103–114.

Asli Çelikyilmaz, Zhaleh Feizollahi, Dilek Hakkani-Tür, and Ruhi Sarikaya. 2014. Resolving referring expressions in conversational dialogs for natural user interfaces. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 2094–2104.

Desai Chen, Nathan Schneider, Dipanjan Das, and Noah A. Smith. 2010. SEMAFOR: frame argument resolution with log-linear models. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 264–267.

Tongfei Chen, Chetan Naik, Hua He, Pushpendre Rastogi, and Lambert Mathias. 2019. Improving long distance slot carryover in spoken dialogue systems. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 96–105. Association for Computational Linguistics.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic frame-semantic parsing. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*, pages 948–956.

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*, pages 4171–4186.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 484–495.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8057–8077.

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

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

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

Heng Ji and Ralph Grishman. 2008. Refining event extraction through cross-document inference. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 254–262.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. 2019. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 3744–3753.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

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*, pages 73–82.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7999–8009.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations*. OpenReview.net.

Trung Minh Nguyen and Thien Huu Nguyen. 2019. One for all: Neural joint modeling of entities and events. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019*, pages 6851–6858.

Timothy J. O'Gorman. 2019. *Bringing Together Computational and Linguistic Models of Implicit Role Interpretation*. Ph.D. thesis, University of Colorado at Boulder.

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

Martha S. Palmer, Deborah A. Dahl, Rebecca J. Schiffman, Lynette Hirschman, Marcia Linebarger, and John Dowding. 1986. Recovering implicit information. In *24th Annual Meeting of the Association for Computational Linguistics*, pages 10–19.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2010. Semeval-2010 task 10: Linking events and their participants in discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 45–50.

Carina Silberer and Anette Frank. 2012. Casting implicit role linking as an anaphora resolution task. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 1–10.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, pages 5998–6008.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. ACE 2005 multilingual training corpus (LDC2006T06). *Philadelphia: Linguistic Data Consortium*.

Ye-Yi Wang, Li Deng, and Alex Acero. 2011. *Semantic Frame Based Spoken Language Understanding*, pages 35–80. Wiley.

Bishan Yang and Tom M. Mitchell. 2016. Joint extraction of events and entities within a document context. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 289–299.

Sheng Zhang, Xutai Ma, Kevin Duh, and Benjamin Van Durme. 2019. AMR parsing as sequence-to-graph transduction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pages 80–94.

A Appendix

Experimental Details We use BERT (BERT-BASE-CASED here) as the encoder for text embedding. The models are setup with $d_{\text{tok}} = d_{\text{span}} = 768$, and are trained using AdamW optimizer (Loshchilov and Hutter, 2019) with learning rate of 3×10^{-5} for 200 epochs, and the tolerance $\epsilon = 1 \times 10^{-8}$. We employ gradient clipping to avoid exploding gradients with maximum gradient norm 5.0. We also use a linear learning rate scheduler to warmup models for the first 200 iterations.

The Transformer encoder has 3 layers with 64 attention heads[9], and its feed-forward neural networks (FFNNs) for computing the argument / trigger representations are set to have the dim of 2,048. For mention representations, we use two-layer FFNNs with hidden size of 768. Note there are two different sets of parameters for constructing trigger representations and argument representations. All non-linearities used in the paper are GELU (Hendrycks and Gimpel, 2016). Dropout with rate 0.2 is applied in each levels in the feed-forward neural network for argument / trigger representation computation, and also in each layer in the Transformer encoder.

For model selection, we pick the best performing model on the dev set and then run it on the test set. Early stopping is used with patience $p = 10$, i.e., if the performance on the dev set did not increase after p epochs, stop training.

In terms of hyperparameter sweep, we perform grid search over a combination of hyperparameters shown in Table 6, and choose the set performed best on the dev set.

Our models are trained on one Nvidia GTX 1080 Ti GPU. For the joint model, the training time is around 30 mins/epoch, and it takes 70 epochs (around 20 hours) to converge on average. For the independent model, it takes 15mins/epoch and converges in 5 epochs (around 50 mins) on average.

Hyperparameter	Range
# Encoder layers	$\{1, 2, 3, 4, 5, 6\}$
# Attention heads	$\{12, 64, 128\}$
Learning rate	$\{1 \times 10^{-5}, 3 \times 10^{-5}, 5 \times 10^{-5}\}$
Warmup steps	$\{0, 100, 200, \cdots, 500, 1000\}$

Table 6: Ranges for hyperparameter sweeps.

[9] According to Chen et al. (2019), increasing the number of attention heads substantially improves the model performance, so we prefer more attention heads over more encoder layers.