Saliency Learning: Teaching the Model Where to Pay Attention*

Reza Ghaeini, Xiaoli Z. Fern, Hamed Shahbazi, Prasad Tadepalli
School of Electrical Engineering and Computer Science, Oregon State University
1148 Kelley Engineering Center, Corvallis, OR 97331-5501, USA
{ghaeinim, xfern, shahbazh, tadepall}@eecs.oregonstate.edu

Abstract

Deep learning has emerged as a compelling solution to many NLP tasks with remarkable performances. However, due to their opacity, such models are hard to interpret and trust. Recent work on explaining deep models has introduced approaches to provide insights toward the model's behavior and predictions, which are helpful for determining the reliability of the model's prediction. However, such methods do not fix and improve the model's reliability. In this paper, we teach our models to make the right prediction for the right reason by providing explanation training signal and ensuring alignment of the models explanation with the ground truth explanation. Our experimental results on multiple tasks and datasets demonstrate the effectiveness of the proposed method, which produces more reliable predictions while delivering better results compared to traditionally trained models.

1 Introduction

It is unfortunate that our data is often plagued by meaningless or even harmful statistical biases. When we train a model on such data, it is possible that the classifier would focus on irrelevant biases to achieve high performance on the biased data. Recent studies demonstrate that deep learning models noticeably suffer from this issue (Agrawal et al., 2016; Wadhwa et al., 2018; Gururangan et al., 2018). Due to the black-box nature of deep models and the high dimensionality of their inherent representations, it is difficult to interpret and trust their behaviour and predictions. Recent work on explanation and interpretation has introduced a few approaches (Simonyan et al., 2013; Ribeiro et al., 2016; Lei et al., 2016; Li et al., 2016, 2017; Ghaeini et al., 2018b; Ribeiro et al., 2018) for explanation. Such methods provide insights toward the model's behaviour, which is helpful for detecting biases in our models. However, they do not correct them. In this work, we investigate how to incorporate explanations into the learning process to ensure that our model not only makes correct predictions but also makes them for the right reason.

Specifically, we propose to train a deep model using both traditional ground truth labels and additional annotations suggesting the desired explanation. The learning is achieved via a novel method called *saliency learning*, which regulates the model's behavior using saliency to ensure that the most critical factors impacting the model's prediction is aligned with the desired explanation.

Our work is closely related to Ross et al. (2017), which also uses the gradient/saliency information to regularize model's behaviour. However, we differ in the following points: 1) Ross et al. (2017) is limited to regularizing model with gradient of model's input. In contrast, we extend this concept to the intermediate layers of deep models. 2) Ross et al. (2017) consider a dimension-level annotation and regularization, while we believe annotation should be word-level. 3) Ross et al. (2017) utilize random annotation for finding different decision boundaries, however, we are looking for gold annotation to obtain a reliable model. 4) We utilize a different formulation and regularization.

We make four main contributions: 1) Proposing a new method for teaching the model where to pay attention. 2) Achieving more reliable predictions while delivering better results than traditionally trained models. 3) Evaluating our method on multiple tasks and datasets to demonstrate its effectiveness and generality. 4) Verifying the sensitivity of the trained model using our method (saliency learning) to the contributory parts of the data.

*Accepted as a short paper at NAACL HLT 2019.

2 Saliency-based Explanation Learning

Our goal is to teach the model where to pay attention to prevent focusing on meaningless statistical biases in the data. In this work, we focus on positive explanations. In other words, we expect the explanation to highlight information that contributes positively towards the label. For example, if a piece of text contains the mention of a particular event, then the explanation will highlight parts of the text indicating the event, not non-existence of some other events. This choice is because positive evidence is more natural for humans to specify.

Formally, each training example is a tuple (X, y, Z), where $X = [X_1, X_2, \ldots, X_n]$ is the input text (length n), y is the ground-truth label, and $Z \in \{0,1\}^n$ is the ground-truth explanation as a binary mask indicating whether each word contributes positive evidence toward the label y.

Recent studies have shown that model's predictions can be explained by looking at the saliency of the inputs (Simonyan et al., 2013; Hechtlinger, 2016; Ross et al., 2017; Li et al., 2016) as well as other internal elements of the model (Ghaeini et al., 2018b). Given an example, for which the model makes a prediction, the saliency of a particular element is computed as the derivative of the model's prediction with respect to that element. Saliency provides clues as to where the model is drawing strong evidence to support its prediction. As such, if we constrain the saliency to be aligned with the desired explanation during learning, our model will be coerced to pay attention to the right evidence.

In computing saliency, we are dealing with high-dimensional data. For example, each word is represented by an embedding of d dimensions. To aggregate the contribution of all dimensions, we consider the sum of the gradients of all dimensions as the overall vector/embedding contribution. For the i-th word, if $Z(i) = 1$, then its vector should have a positive gradient/contribution, otherwise the model would be penalized. To do this, we incorporate a saliency regularization term to the model cost function using hinge loss. Equation 1 describes our cost function evaluated on a single example (X, y, Z).

$$\begin{aligned} &\mathcal{C}(\theta, X, y, Z) = \mathcal{L}(\theta, X, y) \\ &+ \lambda \sum_{i=1}^{n} \max\left(0, -Z_i \sum_{j=1}^{d} \frac{\partial f_\theta(X, y)}{\partial X_{i,j}}\right) \end{aligned} \quad (1)$$

where $\mathcal{L}$ is a traditional model cost function (e.g. cross-entropy), λ is the hyper parameter, f specifies the model with parameter θ, and $\frac{\partial f}{\partial X_{i,j}}$ represents the saliency of the j-th dimension of word X_i. The new term in the $\mathcal{C}$ penalizes negative gradient for marked words in Z (contributory words).

Since $\mathcal{C}$ is differentiable respect to θ, it can be optimized using existing gradient-based optimization methods. It is important to note that while Equation 1 only regularizes the saliency of the input layer, the same principle can be applied to the intermediate layers of the model (Ghaeini et al., 2018b) by considering the intermediate layer as the input for the later layers.

Note that if $Z = 0$ then $\mathcal{C} = \mathcal{L}$. So, in case of lacking proper annotation for a specific sample or sequence, we can simply use 0 as its annotation. This property enables our method to be easily used in semi-supervised or active learning settings.

3 Tasks and Datasets

To teach the model where to pay attention, we need ground-truth explanation annotation Z, which is difficult to come by. As a proof of concept, we modify two well known real tasks (Event Extraction and Cloze-Style Question Answering) to simulate approximate annotations for explanation. Details of the main tasks and datasets could be found in section B of the Appendix. We describe the modified tasks as follows:

1) **Event Extraction:** Given a sentence, the goal is to determine whether the sentence contains an event. Note that event extraction benchmarks contain annotation of event triggers, which we use to build the annotation Z. In particular, the Z value of every word is annotated to be zero unless it belongs to an event trigger. For this task, we consider two well known event extraction datasets, namely ACE 2005 and Rich ERE 2015.

2) **Cloze-Style Question Answering:** Given a sentence and a query with a blank, the goal is to determine whether the sentence contains the correct replacement for the blank. Here, annotation of each word is zero unless it belongs to the gold replacement. For this task, we use two well known

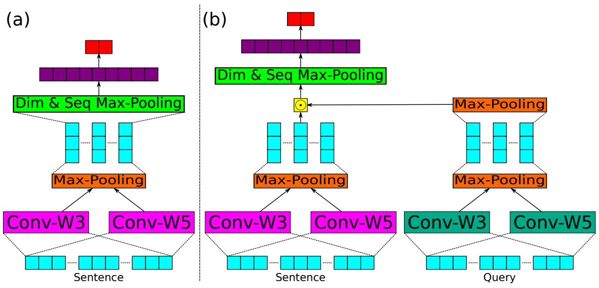

Figure 1: A high-level view of the models used for event extraction (a) and question answering (b).

cloze-style question answering datasets: Children Book Test Named Entity (CBT-NE) and Common Noun (CBT-CN) (Hill et al., 2015).

Here, we only consider the simple binary tasks as a first attempt to examine the effectiveness of our method. However, our method is not restricted to binary tasks. Note that for both tasks if an example is negative, its explanation annotation will be all zero. In other words, for negative examples we have $\mathcal{C} = \mathcal{L}$.

4 Model

We use simple CNN based models to avoid complexity. Figure 1 illustrates the models used in this paper. Both models have a similar structure. The main difference is that QA has two inputs (sentence and query). We first describe the event extraction model followed by the QA. model.

Figure 1 (a) shows the event extraction model. Given a sentence $W = [w_1, \dots, w_n]$ where $w_i \in \mathbb{R}^d$, we first pass the embeddings to two CNNs with feature size of d and window size of 3 and 5. Next we apply max-pooling to both CNN outputs. It will give us the representation $I \in \mathbb{R}^{n \times d}$, which we refer to it as the *intermediate representation*. Then, we apply sequence-wise and dimension-wise max-poolings to I to capture $D_{seq} \in \mathbb{R}^d$ and $D_{dim} \in \mathbb{R}^n$ respectively. D_{dim} will be referred as *decision representation*. Finally we pass the concatenation of D_{seq} and D_{dim} to a feed-forward layer for prediction.

Figure 1 (b) depicts the QA model. The main difference is having *query* as an extra input. To process the query, we use a similar structure as the main model. After CNNs and max-pooling we end up with $Q \in \mathbb{R}^{m \times d}$ where m is the length of query. To obtain a sequence independent vector, we apply another max-pooling to Q resulting in a query representation $q \in \mathbb{R}^d$. We follow a similar approach on the sentence as in event extraction.

Dataset	S.[a]	P.[b]	R.[c]	F1	Acc.[d]
ACE	No	66.0	**77.5**	71.3	74.4
	Yes	**70.1**	76.1	**73.0**	**76.9**
ERE	No	85.0	86.6	85.8	83.1
	Yes	**85.8**	**87.3**	**86.6**	**84.0**
CBT-NE	No	55.6	**76.3**	64.3	75.5
	Yes	**57.2**	74.5	**64.7**	**76.5**
CBT-CN	No	47.4	**39.0**	42.8	77.3
	Yes	**48.3**	38.9	**43.1**	**77.7**

[a]Saliency Learning. [b]Precision. [c]Recall. [d]Accuracy

Table 1: Performance of trained models on multiple datasets using traditional method and saliency learning.

The only difference is that we apply the dot product between the *intermediate representations* and query representation ($I_i = I_i \odot q$).

As mentioned previously, we can apply saliency regularization to different levels of the model. In this paper, we apply saliency regularization on the following three levels: 1) Word embeddings (W). 2) Intermediate representation (I). 3) Decision representation (D_{dim}). Note that the aforementioned levels share the same annotation for training. For training details please refer to Section C of the Appendix.

5 Experiments and Analysis

5.1 Performance

Table 1 shows the performance of the trained models on ACE, ERE, CBT-NE, and CBT-CN datasets using the aforementioned models with and without saliency learning. The results indicate that using saliency learning yields better accuracy and F1 measure on all four datasets. It is interesting to note that saliency learning consistently helps the models to achieve noticeably higher precision without hurting the F1 measure and accuracy. This observation suggests that saliency learning is effective in providing proper guidance for more accurate predictions – Note that here we only have guidance for positive prediction. To verify the statistical significance of the observed performance improvement over traditionally trained models without saliency learning, we conducted the one-sided McNemar's test. The obtained p-values are 0.03, 0.03, 0.0001, and 0.04 for ACE, ERE, CBT-NE, and CBT-CN respectively, indicating that the performance gain by saliency learning

Dataset	S.	W.[a]	I.[b]	D.[c]
ACE	No	61.60	66.05	63.27
	Yes	**99.26**	**77.92**	**65.49**
ERE	No	51.62	56.71	44.37
	Yes	**99.77**	**77.45**	**51.78**
CBT-NE	No	52.32	65.38	68.81
	Yes	**98.17**	**98.34**	**95.56**
CBT-CN	No	47.78	53.68	45.15
	Yes	**99.13**	**98.94**	**97.06**

[a]Word Level Saliency Accuracy.
[b]Intermediate Level Saliency Accuracy.
[c]Decision Level Saliency Accuracy.

Table 2: Saliency accuracies of different layer of our models trained on ACE, ERE, CBT-NE, CBT-CN.

Dataset	S.	TPR_0^a	TPR_1^b	ΔTPR^c
ACE	No	77.5	52.2	32.6
	Yes	76.1	45.0	**40.9**
ERE	No	86.6	73.2	15.4
	Yes	87.3	70.6	**19.1**
CBT-NE	No	76.3	30.2	60.4
	Yes	74.5	28.5	**61.8**
CBT-CN	No	39.0	16.6	57.4
	Yes	38.9	15.4	**60.4**

[a]True Positive Rate (before removal).
[b]TPR after removing the critical word(s).
[c]TPR change rate.

Table 3: True positive rate and true positive rate change of the trained models before and after removing the contributory word(s).

is statistically significant.

5.2 Saliency Accuracy and Visualization

In this section, we examine how well does the saliency of the trained model aligns with the annotation. To this end, we define a metric called *saliency accuracy* (s_{acc}), which measures what percentage of all positive positions of annotation Z indeed obtain a positive gradient. Formally, $s_{acc} = 100\frac{\sum_i \delta(Z_i G_i > 0)}{\sum_i Z_i}$ where G_i is the gradient of unit i and δ is the indicator function.

Table 2 shows the saliency accuracies at different layers of the trained model with and without saliency learning. According to Table 2, our method achieves a much higher saliency accuracy for all datasets indicating that the learning was indeed effective in aligning the model saliency with the annotation. In other words, important words will have positive contributions in the saliency-trained model, and as such, it learns to focus on the right part of the data. This claim can also be verified by visualizing the saliency, which are provided in section D of the Appendix.

5.3 Verification

Up to this point we show that using saliency learning yields noticeably better precision, F1 measure, accuracy, and saliency accuracy. Here, we aim to verify our claim that saliency learning coerces the model to pay more attention to the critical parts. The annotation Z describes the influential words toward the positive labels. Our hypothesis is that *removing such words would cause more impact on saliency-trained models*, since by training they should be more sensitive to these words. We measure the impact as the percentage change of the model's true positive rate. This measure is chosen because negative examples do not have any annotated contributory words, and hence we are particularly interested in how removing contributory words of positive examples would impact the model's true positive rate (TPR).

Table 3 shows the outcome of the aforementioned experiment, where the last column lists the TPR reduction rates. From the table, we see a consistently higher rate of TPR reduction for saliency-trained models compared to traditionally trained models, suggesting that saliency-trained models are more sensitive to the presence of the contributory words and confirming our hypothesis.

It is worth noting that we observe less substantial change to the true positive rate for the event task. This is likely due to the fact that we are using the trigger words as simulated explanations. While trigger words are clearly related to events, there are often other words in the sentence relating to events but not annotated as trigger words.

6 Conclusion

In this paper, we proposed *saliency learning*, a novel approach for teaching a model where to pay attention. We demonstrated the effectiveness of our method on multiple tasks and datasets using simulated explanations. The results show that saliency learning enables us to obtain better precision, F1 measure and accuracy on these tasks and datasets. Further, it produces models whose saliency is more properly aligned with the desired explanation. In other words, *saliency learning* gives us more reliable predictions while delivering better performance than traditionally trained models. Finally, our verification experiments illustrate that the saliency trained models show higher sensitivity to the removal of contributory words in a positive example. For future work, we will extend our study to examine saliency learning on NLP tasks in an active learning setting where real explanations are requested and provided by human.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1955–1960.

Yubo Chen, Liheng Xu, Kang Liu, Daojian Zeng, and Jun Zhao. 2015. Event extraction via dynamic multi-pooling convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 167–176.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. 2017. Attention-over-attention neural networks for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 593–602.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. 2017. Gated-attention readers for text comprehension. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1832–1846.

Reza Ghaeini, Xiaoli Z. Fern, Liang Huang, and Prasad Tadepalli. 2016. Event nugget detection with forward-backward recurrent neural networks. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*.

Reza Ghaeini, Xiaoli Z. Fern, Hamed Shahbazi, and Prasad Tadepalli. 2018a. Dependent gated reading for cloze-style question answering. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3330–3345.

Reza Ghaeini, Xiaoli Z. Fern, and Prasad Tadepalli. 2018b. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4952–4957.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 107–112.

Yotam Hechtlinger. 2016. Interpretation of prediction models using the input gradient. *CoRR*, abs/1611.07634.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *CoRR*, abs/1511.02301.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Tao Lei, Regina Barzilay, and Tommi S. Jaakkola. 2016. Rationalizing neural predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 107–117.

Jiwei Li, Xinlei Chen, Eduard H. Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 681–691.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220.

John Walker Orr, Prasad Tadepalli, and Xiaoli Z. Fern. 2018. Event detection with neural networks: A rigorous empirical evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 999–1004.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1527–1535.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2662–2670.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034.

Adam Trischler, Zheng Ye, Xingdi Yuan, Philip Bachman, Alessandro Sordoni, and Kaheer Suleman. 2016. Natural language comprehension with the epireader. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 128–137.

Soumya Wadhwa, Varsha Embar, Matthias Grabmair, and Eric Nyberg. 2018. Towards inference-oriented reading comprehension: Parallelqa. *CoRR*, abs/1805.03830.

A Background: Saliency

The concept of saliency was first introduced in vision for visualizing the spatial support on an image for particular object class (Simonyan et al., 2013). Considering a deep model prediction as a differentiable model f parameterized by θ with input $X \in \mathbb{R}^{n \times d}$. Such model could be describe using the Taylor series as follow:

$$f(x) = f(a) + f^{'}(a)(x-a) + \frac{f^{''}(a)}{2!}(x-a)^2 + \dots \tag{2}$$

By approximating that the deep model is a linear function, we could just use the first order Taylor expansion.

$$f(x) \approx f^{'}(a)x + b \tag{3}$$

According to Equation 3, the first derivative of model's prediction with respect to the input ($f^{'}(a)$ or $\frac{\partial f}{\partial x}|_{x=a}$) serves as the description of model's behaviour near the input. To make it more clear, bigger derivative/gradient indicates more impact and contribution toward model's prediction. Consequently, the large-magnitude derivative values determine units of input that would greatly affect $f(x)$ if changed.

B Task and Dataset

Here, we first describe the main and real Event Extraction and Close-Style Question Answering tasks (before our modification). Next, we provide data statistics of the modified version of ACE, ERE, CBT-NE, and CBT-CN datasets in Table 4.

- **Event Extraction:** Given a set of ontologized event types (e.g. Movement, Transaction, Conflict, etc.), the goal of event extraction is to identify the mentions of different events along their types from natural texts (Chen et al., 2015; Ghaeini et al., 2016; Orr et al., 2018).

- **Cloze-Style Question Answering:** Documents in CBT consist of 20 contiguous sentences from the body of a popular children book, and queries are formed by replacing a token from the 21^{st} sentence with a blank. Given a document, a query, and a set of candidates, the goal is to find the correct replacement for blank in the query among the given candidates. To avoid having too many negative examples in our modified datasets, we only consider the sentences that contains at least one candidate. To be more clear, each sample from the CBT dataset is split to at most 20 samples – each sentence of the main sample as long as it contains one of the candidates (Trischler et al., 2016; Kadlec et al., 2016; Cui et al., 2017; Dhingra et al., 2017; Ghaeini et al., 2018a).

C Training

All hyper-parameters are tuned based on the development set. We use pre-trained $300-D$ Glove $840B$ vectors (Pennington et al., 2014) to initialize our word embedding vectors. All hidden states and feature sizes are 300 dimensions ($d = 300$). The weights are learned by minimizing the

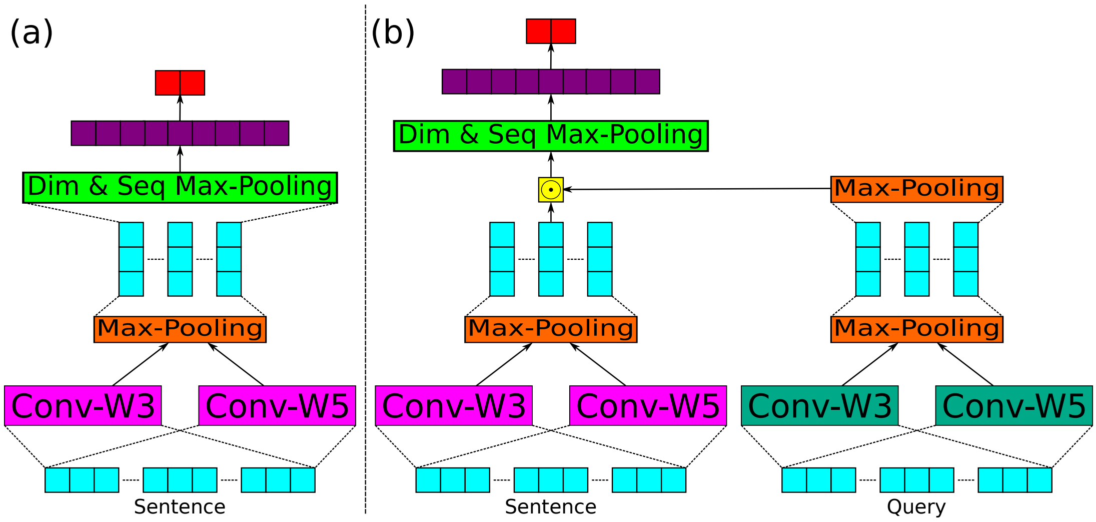

Figure 2: A high-level view of the models used for event extraction(a) and question answering (b).

Dataset	Sample Count			
	Train		Test	
	P.[a]	N.[b]	P.	N.
ACE	3.2K	15K	293	421
ERE	3.1K	4K	2.7K	1.91K
CBT-NE	359K	1.82M	8.8K	41.1K
CBT-CN	256K	2.16M	5.5K	44.4K

[a] Positive Sample Count
[b] Negative Sample Count

Table 4: Dataset statistics of the modified tasks and datasets.

cost function on the training data via Adam optimizer. The initial learning rate is 0.0001 and $\lambda = 0.5, 0.7, 0.4$, and 0.35 for ACE, ERE, CBT-NE, and CBT-CN respectively. To avoid overfitting, we use dropout with the rate of 0.5 for regularization, which is applied to all feedforward connections. During training, the word embeddings are updated to learn effective representations for each task and dataset. We use a fairly small batch size of 32 to provide more exploration power to the model.

D Saliency Visualization

In this section, we empirically analyze the traditionally trained (Baseline Model) and saliency-based trained model (saliency-based Model) behaviours by observing the saliency of 23 positive samples from ACE and ERE datasets. Tables 5 and 6 show the top 6 salient words (words with highest saliency/gradient) of a positive sample from ACE or ERE dataset along with its contributory words (Z), baseline model prediction (P_B), and saliency-based model prediction (P_S). Darker red color indicates more salient words. Our observations could be divided to six categories as follow:

- Samples 1-7: Both models correctly predict 1 for these samples. Saliency-based model successfully pays attention to the expected meaningful words while Baseline model pays attention to mostly irrelevant ones.

- Samples 8-11: Both models correctly predict 1 and pays attention to the contributory words. Yet, we observe lower saliency for important words and higher attention for irrelevant ones.

- Samples 12-14: Here, Baseline model fails to pay attention to the contributory words and predicts 0 while saliency-based one successfully pays attention to them and predicts 1.

- Samples 15-18: Although the models have high saliency for the contributory words, still they could not correctly disambiguate these samples. This observation suggests that having high saliency for important words does not guarantee positive prediction. High saliency for these words indicate their positive contribution toward the positive prediction but still the model might consider higher probability for negative prediction.

- Samples 19-21: Here, only Baseline model could correctly predict 1. However, Baseline model does not pay attention to the contributory words. In other words, the explanation does not support the prediction (unreliable).

- Samples 22-23: Not always the saliency-based model could pay enough attention to the contributory words. In these examples, baseline model has high saliency for contributory words. It worth noting that when saliency-based model does not have high saliency for contributory words, it does not predict positive prediction. Such observation could suggest that saliency-based model predictions are more reliable. The aforementioned observation is also verified by consistently obtaining noticeably higher precision (Section 5.1 and Table 1 in the main paper).

id	Baseline Model	Saliency-based Model	Z	P_B	P_S
1	The judge at Hassan's extradition hearing said that he found the French handwriting report very problematic, very confusing, and with suspect conclusions.	The judge at Hassan 's extradition hearing said that he found the French handwriting report very problematic, very confusing, and with suspect conclusions.	extradition hearing said	1	1
2	Solana said the EU would help in the humanitarian crisis expected to follow an attack on Iraq.	Solana said the EU would help in the humanitarian crisis expected to follow an attack on Iraq .	attack	1	1
3	The trial will start on March 13 , the court said .	The trial will start on March 13 , the court said.	trial	1	1
4	India 's has been reeling under a heatwave since mid-May which has killed 1,403 people.	India 's has been reeling under a heatwave since mid-May which has killed 1,403 people .	killed	1	1
5	Retired General Electric Co. Chairman Jack Welch is seeking work-related documents of his estranged wife in his high-stakes divorce case .	Retired General Electric Co. Chairman Jack Welch is seeking work-related documents of his estranged wife in his high-stakes divorce case .	Retired divorce	1	1
6	The following year, he was acquitted in the Guatemala case, but the U.S. continued to push for his prosecution.	The following year, he was acquitted in the Guatemala case , but the U.S. continued to push for his prosecution .	acquitted case	1	1
7	In 2011, a Spanish National Court judge issued arrest warrants for 20 men , including Montano,suspected of participating in the slaying of the priests.	In 2011, a Spanish National Court judge issued arrest warrants for 20 men, including Montano,suspected of participating in the slaying of the priests.	issued slaying arrest	1	1
8	Slobodan Milosevic's wife will go on trial next week on charges of mismanaging state property during the former president's rule, a court said Thursday.	Slobodan Milosevic's wife will go on trial next week on charges of mismanaging state property during the former president 's rule, a court said Thursday .	trial charges former	1	1
9	Iraqis mostly fought back with small arms, pistols, machine guns and rocket-propelled grenades .	Iraqis mostly fought back with small arms, pistols, machine guns and rocket-propelled grenades.	fought	1	1
10	But the Saint Petersburg summit ended without any formal declaration on Iraq .	But the Saint Petersburg summit ended without any formal declaration on Iraq .	summit	1	1

Table 5: Top 6 salient tokens visualization of samples in ACE and ERE for baseline and saliency-based models.

id	Baseline Model	Saliency-based Model	Z	P_B	P_S
11	He will then stay on for a regional summit before heading to Saint Petersburg for celebrations marking the 300th anniversary of the city's founding .	He will then stay on for a regional summit before heading to Saint Petersburg for celebrations marking the 300th anniversary of the city's founding.	heading summit	1	1
12	From greatest moment of his life to divorce in 3 years or less.	From greatest moment of his life to divorce in 3 years or less.	divorce	0	1
13	The state s execution record has often been criticized .	The state s execution record has often been criticized .	execution	0	1
14	The student, who was 18 at the time of the alleged sexual relationship, testified under a pseudonym .	The student, who was 18 at the time of the alleged sexual relationship , testified under a pseudonym.	testified	0	1
15	U.S. aircraft bombed Iraqi tanks holding bridges close to the city .	U.S. aircraft bombed Iraqi tanks holding bridges close to the city.	bombed	0	0
16	However , no blasphemy convict has ever been executed in the country .	However, no blasphemy convict has ever been executed in the country .	executed	0	0
17	Gul 's resignation had been long expected .	Gul 's resignation had been long expected .	resignation	0	0
18	aside from purchasing alcohol, what rights don't 18 year olds have?	aside from purchasing alcohol , what rights don't 18 year olds have?	purchasing	0	0
19	He also ordered him to have no contact with Shannon Molden.	He also ordered him to have no contact with Shannon Molden .	ordered contact	1	0
20	This means your account is once again active and operational, Riao wrote Colombia Reports.	This means your account is once again active and operational , Riao wrote Colombia Reports .	wrote	1	0
21	I am a Christian as is my ex husband yet we are divorced.	I am a Christian as is my ex husband yet we are divorced .	divorced ex	1	0
22	Taylor acknowledged in his testimony that he ran up toward the pulpit with a large group and followed the men outside.	Taylor acknowledged in his testimony that he ran up toward the pulpit with a large group and followed the men outside.	testimony followed ran	1	0
23	The note admonished Jasper Molden , and his then-fiance, Shannon Molden .	The note admonished Jasper Molden , and his then-fiance , Shannon Molden.	note	0	0

Table 6: Top 6 salient tokens visualization of samples in ACE and ERE for baseline and saliency-based models.