A Dual-Attention Network for Joint Named Entity Recognition and Sentence Classification of Adverse Drug Events

Susmitha Wunnava
Worcester Polytechnic Institute
100 Institute Rd
Worcester, MA 01609
swunnava@wpi.edu

Xiao Qin
IBM Research - Almaden
650 Harry Road
San Jose CA 95120
xiao.qin@ibm.com

Tabassum Kakar
Worcester Polytechnic Institute
100 Institute Rd
Worcester, MA 01609
tkakar@wpi.edu

Xiangnan Kong
Worcester Polytechnic Institute
100 Institute Rd
Worcester, MA 01609
xkong@wpi.edu

Elke A. Rundensteiner
Worcester Polytechnic Institute
100 Institute Rd
Worcester, MA 01609
rundenst@wpi.edu

Abstract

An adverse drug event (ADE) is an injury resulting from medical intervention related to a drug. ADE detection from text can be either fine-grained (ADE entity recognition) or coarse-grained (ADE assertive sentence classification), with limited efforts leveraging inter-dependencies among these two granularities. We instead design a multi-grained joint deep network model MGADE to concurrently solve both ADE tasks MGADE takes advantage of their symbiotic relationship, with a transfer of knowledge between the two levels of granularity. Our dual-attention mechanism constructs multiple distinct representations of a sentence that capture both task-specific and semantic information in the sentence, providing stronger emphasis on the key elements essential for sentence classification. Our model improves state-of-art F1-score for both tasks: (i) entity recognition of ADE words (12.5% increase) and (ii) ADE sentence classification (13.6% increase) on MADE 1.0 benchmark of EHR notes.

1 Introduction

Background. Adverse drug events (ADEs), injuries resulting from medical intervention, are a leading cause of death in the United States and cost around \$30˜\$130 billion every year (Donaldson et al., 2000). Early detection of ADE incidents aids in the timely assessment, mitigation and prevention of future occurrences of ADEs. Natural Language Processing techniques have been recognized as instrumental in identifying ADEs and related information from unstructured text fields of spontaneous reports and electronic health records (EHRs) and thus in improving drug safety monitoring and pharmacovigilance (Harpaz et al., 2014).

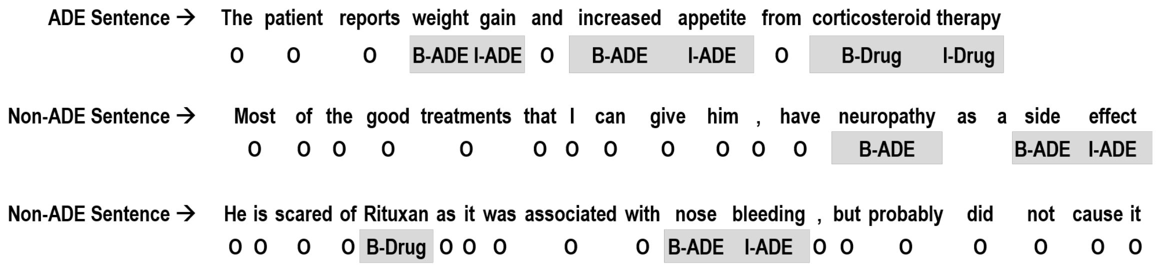
Fine-grained ADE detection identifies named ADE entities at the word-level, while *coarse-grained ADE detection* (also ADE assertive text classification) identifies complete sentences describing drug-related adverse effects. (Gurulingappa et al., 2011)'s system for identification of ADE assertive sentences in medical case reports targets the important application of detecting under-reported and under-documented adverse drug effects. Lastly, *multi-grained ADE detection* identifies ADE information at multiple levels of granularity, namely, both entity and sentence level.

As example, Figure 1 displays ADE and non-ADE sentences. The first is an ADE sentence where the mentions of Drugname and ADE entities have the appropriate relationship with each other. Second and third sentences show that the mention of an ADE entity by itself is not sufficient to assert a drug-related adverse side effect.

Recently, deep learning-based sequence approaches have shown some promise in extracting fine-grained ADEs and related named entities from text (Liu et al., 2019). However, the prevalence of entity-type ambiguity remains a major hurdle, such as, distinguishing between *Indication entities* as the reason for taking a drug versus *ADE entities* as unintended outcomes of taking a drug. Coarse-grained sentence-level detection performs well in identifying ADE descriptive sentences, but is not equipped to detect fine-grained information such as words associated with ADE related named entities. Unfortunately, when the interaction between these two extraction tasks is ignored, we miss the opportunity of the transfer of knowledge between the ADE entity and sentence prediction tasks.

Attention-based neural network models have been shown to be effective for text classification

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 3414–3423
November 16 - 20, 2020. ©2020 Association for Computational Linguistics



ADE Sentence → The patient reports weight gain and increased appetite from corticosteroid therapy
O O O B-ADE I-ADE O B-ADE I-ADE O B-Drug I-Drug

Non-ADE Sentence → Most of the good treatments that I can give him , have neuropathy as a side effect
O O O O O O O O O O O O B-ADE B-ADE I-ADE

Non-ADE Sentence → He is scared of Rituxan as it was associated with nose bleeding , but probably did not cause it
O O O O B-Drug O O O O O B-ADE I-ADE O O O O O O O

Figure 1: Each sentence is classified as ADE sentence (binary yes/no). Each word is labeled using beginning of an entity (B-...) vs inside an entity (I-...) for ADE related named entities (multiple classes). O denotes no entity tag.

tasks (Luong et al., 2015; Bahdanau et al., 2014) from alignment attention in translation (Liu et al., 2016) to supervising attention in binary text classification (Rei and Søgaard, 2019). Previous approaches typically apply only a single round of attention focusing on simple semantic information In our ADE detection task, instead, key elements of the sentence can be linked to multiple categories of task-specific semantic information of the named entities (ADE, Drug, Indication, Severity, Dose etc.). Thus, single attention is insufficient in exploring this multi-aspect information and consequently risks losing important cues.

Proposed Approach. In our work, we tackle the above shortcomings by designing a dual-attention based neural network model for multi-grained joint learning, called MGADE, that jointly identifies both ADE entities and ADE assertive sentences. The design of MGADE is inspired by multi-task Recurrent Neural Network architectures for jointly learning to label tokens and sentences in a binary classification setting (Rei and Søgaard, 2019). In addition, our model makes use of a supervised self-attention mechanism based on entity-level predictions to guide the attention function – aiding it in tackling the above entity-type ambiguity problem. We also introduce novel strategies of constructing multiple complementary sentence-level representations to enhance the performance of sentence classification.

Our key contributions include:

1. *Joint Model.* We jointly model ADE entity recognition as a multi-class sequence tagging problem and ADE assertive text classification as binary classification. Our model leverages the mutually beneficial relationships between these two tasks, e.g., ADE sentence classification can influence ADE entity recognition by identifying clues that contribute to ADE assertiveness of the sentence and match them to ADE entities.
2. *Dual-Attention.* Our novel method for generating and pooling multiple attention mechanisms produces informative sentence-level representations. Our dual-attention mechanisms based on word-level entity predictions construct multiple representations of the same sentence. The dual-attention weighted sentence-level representations capture both task-specific and semantic information in a sentence, providing stronger emphasis on key elements essential for sentence classification.
3. *Label-Awareness.* We introduce an augmented sentence-level representation comprised of predicted entity labels which adds label-context to the proposed dual-attention sentence-level representation for better capturing the word-level label distribution and word dependencies within the sentence. This further boosts the performance of the sentence classification task.
4. *Model Evaluation.* We compare our joint model with state-of-art methods for the ADE entity recognition and ADE sentence classification tasks. Experiments on MADE1.0 benchmark of EHR notes demonstrate that our MGADE model drives up the F1-score for both tasks significantly: (i) entity recognition of ADE words by 12.5% and by 23.5% and (ii) ADE sentence classification by 13.6% and by 23.0%, compared to state-of-art single task and joint-task models, respectively.

2 Related Work

Fine-grained ADE Detection. Jagannatha and Yu (2016b) have employed a bidirectional LSTM-CRF model to label named entities from electronic health records of cancer patients. Pandey et al. (2017) proposed a bidirectional recurrent neural network with attention to extract ADRs and classify the relationship between entities from Medline abstracts and EHR datasets. Wunnava et al. (2019) presented a three-layer deep learning architecture for identifying named entities from EHRs, consisting of a Bi-LSTM layer for character-level encoding, a Bi-LSTM layer for word-level encoding, and a CRF layer for structured prediction.

Coarse-grained ADE Detection. Huynh et al. (2016) applies Convolutional Neural Networks using pre-trained word embeddings to detect sentences describing ADEs. Tafti et al. (2017) utilized a feed-forward ANN to discover ADE sentences on PubMed Central data and social media. Dev et al. (2017) developed a binary document classifier using logistic regression, random forests and LSTMs to classify an AE case as serious vs. non-serious.

Multi-grained ADE Detection. Zhang et al. (2018) developed a multi-task learning model that combines entity recognition with document classification to extract the adverse event from a case narrative and classify the case as serious or non-serious. However, they fall short in tackling our problem. Not only do their targeted labels not fall into the drug-related adverse side effects category in which a causal relationship is suspected and required, but their attention model is only simple self-attention. As consequence, MGADE outperforms their model by 23.5% in F1 score for entity recognition and 23.0% for assertive text classification as seen in Section 4.

3 The Proposed Model: MGADE

3.1 Task Definition

In the ADE and medication related information detection task, the entities are *ADE, Drugname, Dose, Duration, Frequency, Indication, Route, Severity* and *Other Signs & Symptoms*. The no-entity tag is O. Because some entities (like weight gain) can have multiple words, we work with a BIO tagging scheme to distinguish between beginning (tag B-...) versus inside of an entity (tag I-...). The notation we use is given in Fig 2. Given a sentence (a sequence of words), task one is the multi-class classification of ADE and medication related named entities in the text sequence, i.e., entity recognition. Task two is the binary classification of a sentence as ADE assertive text. The overall goal is to minimize the weighted sum of entity recognition loss and sentence classification loss.

3.2 Input Embedding Layer

The input of this layer is a sentence represented by a sequence of words $S = \langle w_1, w_2, ..., w_N \rangle$, where N is sentence length. The words are first broken into individual characters and character-level representations which capture the morphology of a word computed with a bidirectional-LSTM over the sequence of characters in the input words. We employ the pre-trained word vector, GloVe (Pennington et al., 2014), to obtain a fixed word embedding of each word. A consolidated dense embedding, comprised of pre-trained word embedding concatenated with a learned character-level representation, is used to represent a word. The output of this layer is $X = [x_1, x_2, ..., x_N]$.

3.3 Contextual Layer

LSTM is a type of recurrent neural network that effectively captures long-distance sequence information and the interaction between adjacent words (Hochreiter and Schmidhuber, 1997). The word representations x_t are given as input to two separate LSTM networks (Bi-LSTM) that scan the sequence forward and backward, respectively. The hidden states learned by the forward and backward LSTMs are denoted as $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$, respectively.

$$\overrightarrow{h}_t = LSTM\left(x_t, \overrightarrow{h}_{t-1}\right) \tag{1}$$

$$\overleftarrow{h}_t = LSTM\left(x_t, \overleftarrow{h}_{t+1}\right) \tag{2}$$

The output of this layer is a sequence of hidden states $H = [h_1, h_2, ..., h_N]$, where h_t is a concatenation of $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$. This way, the hidden state h_t of a word encodes information about the t^{th} word and its context:

$$h_t = \left[\overrightarrow{h}_t; \overleftarrow{h}_t\right] \tag{3}$$

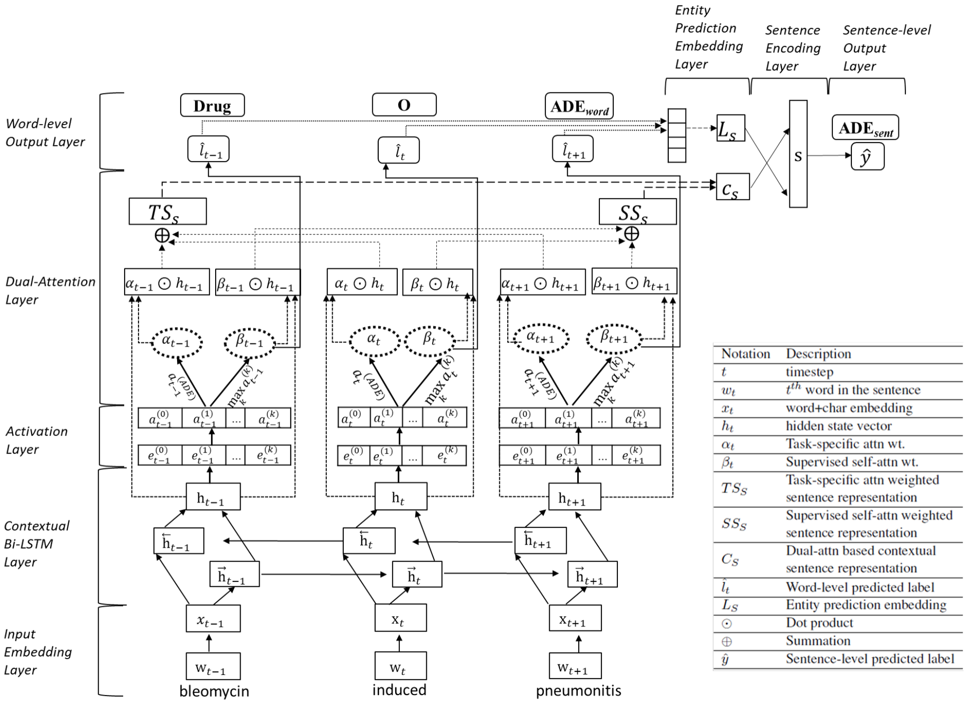
3.4 Word-level (NER) Output Layer

The hidden states h_t are passed through a non-linear layer and then with the softmax activation function to k output nodes, where k denotes the number of entity-types (classes). Entity-type labels are the named entities in the BIO format. Each output node belongs to some entity-type and outputs a score for that entity-type. The output of the softmax function is a categorical probability distribution, where output probabilities of each class is between 0 and 1, and the total sum of all output probabilities is equal to 1.

$$a_t^{(i)} = \frac{\exp\left(e_t^{(i)}\right)}{\sum_{j=1}^{k} \exp\left(e_t^{(j)}\right)} \tag{4}$$

Data is classified into a entity-type that has the highest probability value.

$$\hat{a}_t = \max_{i \in \{1,2,...,k\}} a_t^{(i)} \tag{5}$$



Notation	Description
t	timestep
w_t	t^{th} word in the sentence
x_t	word+char embedding
h_t	hidden state vector
α_t	Task-specific attn wt.
β_t	Supervised self-attn wt.
TS_S	Task-specific attn weighted sentence representation
SS_S	Supervised self-attn weighted sentence representation
C_S	Dual-attn based contextual sentence representation
$\hat{l}_t$	Word-level predicted label
L_S	Entity prediction embedding
$\odot$	Dot product
$\oplus$	Summation
$\hat{y}$	Sentence-level predicted label

Figure 2: The architecture of the proposed **M**ulti-**G**rained **ADE** Detection Network (MGADE)

3.5 Dual-Attention Layer

The purpose of the attention mechanism in the sentence classification task is to select important words in different contexts to build informative sentence representations. Different words have different importance for ADE sentence classification task. For instance, key elements (words/phrases) in the ADE detection task are linked to multiple aspects of semantic information associated with the named entity categories - *ADE, Drugname, Severity, Dose, Duration, Indication... etc*. It is necessary to assign the weight for each word according to its contribution to the ADE sentence classification task.

Moreover, certain named entities are task-specific and are considered essential for ADE sentence classification. There exists a direct correspondence between such task-specific named entities and the sentence. Hence, we anticipate that there would be at least one word of the same label as the sentence-level label. For instance, a sentence that is labeled as an ADE sentence has a corresponding ADE entity word. Although other named entity words detect important information and contribute to the ADE sentence-level classification task, a stronger focus should be on task-specific ADE words indicative of the ADE sentence core message. A single attention distribution tends to be insufficient to explore the multi-aspect information and consequently may risk losing important cues (Wang et al., 2017).

We address this challenge by generating and using multiple attention distributions that offer additional opportunities to extract relevant semantic information. This way, we focus on different aspects of an ADE sentence to create a more informative representation. For this, we introduce a novel dual-attention mechanism, which in addition to selecting the important semantic areas in the sentence (henceforth referred as supervised self-attention (Bahdanau et al., 2014; Yang et al., 2016; Rei and Søgaard, 2019)), it also provides stronger emphasis on task-specific semantic aspect areas (henceforth referred as task-specific attention). The task-specific attention promotes the words important to the ADE sentence-classification task and reduces the noise introduced by words which are less important for the task.

Similar to (Rei and Søgaard, 2019; Yang et al., 2016), we use a self-attention mechanism where, based on softmax probabilities and normalization, attention-weights are extracted from word-level prediction scores. The difference between the two attention mechanism is that the supervised self-attention recognizes word-level prediction scores of all named entities while the task-specific attention recognizes word-level prediction scores w.r.t only selective named entities (one which correspond to the ADE sentence and ignores other named entities). Specifically, the weights of the supervised self-attention and task-specific attention are calculated as follows:

Word-level prediction w.r.t the task-specific named entity (i.e.,) ***ADE***:

$$a_t^{(\boldsymbol{ADE}_{entity})} = \frac{\exp\left(e_t^{(\boldsymbol{ADE}_{entity})}\right)}{\sum_{j=1}^{k}\exp\left(e_t^{(j)}\right)} \quad (6)$$

Task-specific Attention Weight, normalized to sum up to 1 over all values in the sentence, is:

$$\alpha_t = \frac{a_t^{(\boldsymbol{ADE}_{entity})}}{\sum_{n=1}^{N}\left(a_n^{(\boldsymbol{ADE}_{entity})}\right)} \quad (7)$$

Supervised Self-Attention Weight, normalized to sum up to 1 over all values in the sentence:

$$\beta_t = \frac{\hat{a}_t}{\sum_{n=1}^{N}\hat{a}_n} \quad (8)$$

Fig 3 shows the examples of the supervised self-attention and task-specific attention distributions generated from our attention layer. The color depth expresses the degree of importance of the weight in attention vector. As depicted in Fig. 3, the task-specific attention emphasizes more on the parts relevant to the ADE sentence classification task.

Attention-based Sentence Representations. To generate informative and more accurate sentence representations, we construct two different sentence representations as a weighted sum of the context-conditioned hidden states using the task-specific attention weight α_t and supervised self-attention weight β_t, respectively.

1. Task-specific attention weighted sentence rep.:

$$TS_S = \sum_{t=1}^{N}\alpha_t h_t \quad (9)$$

2. Supervised self-attention weighted sentence rep.:

$$SS_S = \sum_{t=1}^{N}\beta_t h_t \quad (10)$$

Attention Pooling A combination of multiple sentence representations obtained from focusing on different aspects captures the overall contextual semantic information about a sentence. The two attention-based representations are concatenated to form a dual-attention contextual sentence representation:

$$C_S = [TS_S; SS_S] \quad (11)$$

3.6 Entity Prediction Embedding Layer

ADE detection is a challenging task. Understanding the co-occurrence of named entities (labels) is essential for ADE sentence classification. Although we implicitly capture long-range label dependencies with Bi-LSTM in the contextual layer, and make even more informative sentence-level representations with the help of the dual-attention layer, explicitly integrating information on the label-distribution in a sentence is further helpful to understand the label co-occurrence structure and dependencies in the sentence. The idea is to further improve the performance of ADE sentence classification task by learning the output word-level label knowledge. For a better representing of the word-level label distribution and to capture potential label dependencies within each sentence, we propose Entity Prediction Embedding (EPE), a sentence-level vector representation of entity labels predicted at the word-level output layer (Sec. 3.4).

$$\hat{l}_t = \underset{i\in\{0,1,2,\ldots,k\}}{\arg\max}\ a_t^{(i)} \quad (12)$$

$$L_S = [v_0, v_1, v_2, ..., v_k]\,; v_i \in \{0, 1\} \quad (13)$$

3.7 Sentence Encoding Layer

A final sentence representation that captures the overall contextual semantic information and label dependencies within the sentence is constructed by combining the dual-attention weighted sentence representation and Entity Prediction Embedding, respectively.

$$\mathbf{S} = [\mathbf{C}_S; \mathbf{L}_S] \quad (14)$$

3.8 Sentence Classification Output Layer

Finally, we apply a fully connected function and use sigmoid activation to output the sentence prediction score.

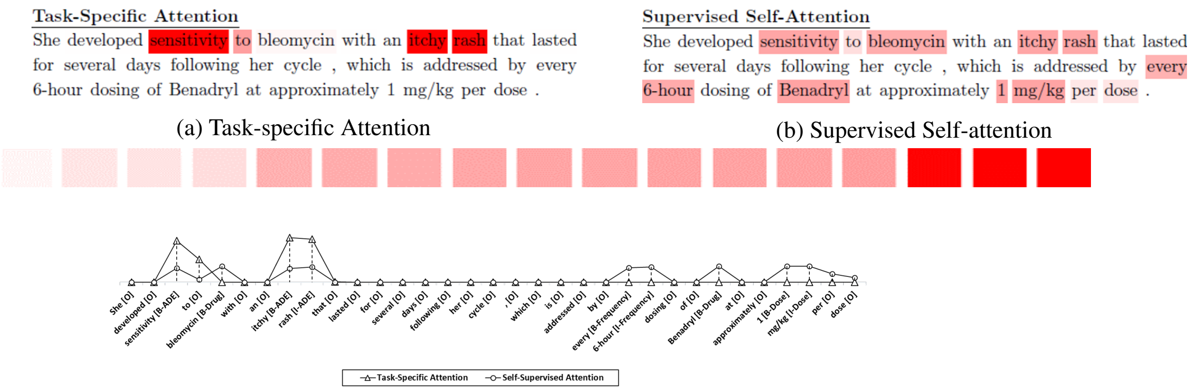
$$\hat{y}^{sentence} = p\left(y^{(j=1)} \mid S\right) \quad (15)$$

3.9 Optimization objective

The objective is to minimize the mean squared error between the predicted sentence-level score $\hat{y}^{(sentence)}$ and the gold-standard sentence label $y^{(sentence)}$ across all m sentences:

$$L_{sentence} = \sum_m \left(y^{(m)} - \hat{y}^{(m)}\right)^2 \quad (16)$$

The objective is to minimize the cross-entropy loss between the predicted word-level probability



Task-Specific Attention

She developed sensitivity to bleomycin with an itchy rash that lasted for several days following her cycle , which is addressed by every 6-hour dosing of Benadryl at approximately 1 mg/kg per dose .

(a) Task-specific Attention

Supervised Self-Attention

She developed sensitivity to bleomycin with an itchy rash that lasted for several days following her cycle , which is addressed by every 6-hour dosing of Benadryl at approximately 1 mg/kg per dose .

(b) Supervised Self-attention

(c) Distribution of attention weights.

Figure 3: Attention Visualizations: Highlighted words indicate attended words. Stronger color denotes higher focus of attention. (a) Task-specific attention: Recognizes task-specific semantic aspect areas of sentence, with focus on ADE entity words essential for ADE sentence classification task. (b) Supervised Self-attention: Recognizes all important areas in the sentence. (c) Distribution of Task-specific attention and Supervised Self-attention weights.

score $\hat{y}^{(entity)}$ and the gold-standard sentence label $y^{(entity)}$ across all N words in the sentence:

$$L_{word} = -\sum_{m}\sum_{t=1}^{N}\sum_{i=1}^{k}\left[a_{ti}^{(m)} log\left(\hat{a}_{ti}^{(m)}\right)\right] \quad (17)$$

Similar to (Rei and Søgaard, 2019), we also add another loss function for joining the sentence-level and word-level objectives that encourages the model to optimize for two conditions on the ADE sentence (i) an ADE sentence must have at least one ADE entity word, and (ii) ADE sentence must have at least one word that is either non-ADE entity or a no-entity word.

$$L_{attn} = \sum_{m}\left(\min\left(\hat{a}_{t,ADE}^{(m)}\right) - 0\right)^2 + \sum_{m}\left(\max\left(\hat{a}_{t,ADE}^{(m)}\right) - y^{(m)}\right)^2 \quad (18)$$

We combine different objective functions using weighting parameters to allow us to control the importance of each objective. The final objective that we minimize during training is then:

$$L = \lambda_{sent} \cdot L_{sent} + \lambda_{word} \cdot L_{word} + \lambda_{attn} \cdot L_{attn} \quad (19)$$

By using word-level entity predictions as attention weights for composing sentence-level representations, we explicitly connect the predictions at both levels of granularity. When both objectives work in tandem, they help improve the performance of one another. In our joint model, we give equal importance to both tasks and set $\lambda_{word} = \lambda_{sentence} = 1$.

4 Experimental Study

4.1 Data Set

MADE1.0 NLP challenge for detecting medication and ADE related information from EHR (Jagannatha and Yu, 2016a) used 1089 de-identified EHR notes from 21 cancer patients (Training: 876 notes, Testing: 213 notes). The annotation statistics of the corpus are provided (Jagannatha et al., 2019).

Named Entity Labels. The notes are annotated with several categories of medication information. *Adverse Drug Event (ADE), Drugname, Indication* and *Other Sign Symptom and Diseases (OtherSSD)* are specified as medical events that contribute to a change in a patient's medical status. *Severity, Route, Frequency, Duration* and *Dosage* specified as attributes describe important properties about the medical events. *Severity* denotes the severity of a disease or symptom. *Route, Frequency, Duration* and *Dosage* as attributes of *Drugname* label the medication method, frequency of dosage, duration of dosage, and the dosage quantity, respectively.

Sentence Labels. MADE 1.0 text has each word manually annotated with ADE or medication related entity types. For words that belong to the ADE entity type, an additional relation annotation denotes if the ADE entity is an adverse side effect of the prescription of the *Drugname* entity. Since MADE 1.0 dataset does not have sentence-level annotations, we use the relation annotation with the word annotation to assign each sentence a label as ADE or nonADE. In this work, the relation labels are used only to assign the sentence labels, but they are not used in the supervised learning process.

4.2 Hyper-parameter Settings

The model operates on tokenized sentences. Tokens were lower-cased, while the character-level component receives input with the original capitalization to learn the morphological features of each word. As input, the pre-trained publicly available Glove word embeddings of size 300 (Pennington

She (0.0) was (0.0) on (0.0) elavil (0.0) before (0.0)
and (0.0) has (0.0) been (0.0) taking (0.0) it (0.0) but (0.0)
it (0.0) dries (0.66) her (0.27) out (0.07)

(a) Single Task-specific Attention

She (0.0) was (0.0) on (0.0) elavil (0.0) before (0.0)
and (0.0) has (0.0) been (0.0) taking (0.0) it (0.0) but (0.0)
it (0.0) dries (0.43) her (0.25) out (0.32)

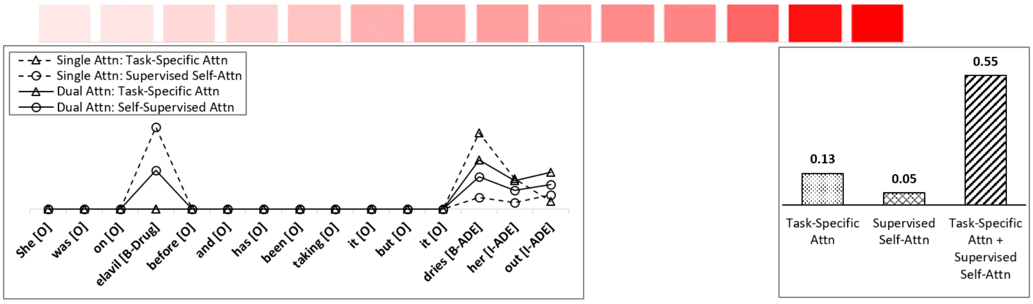
(b) Dual Task-specific attention

She (0.0) was (0.0) on (0.0) elavil (0.71) before (0.0)
and (0.0) has (0.0) been (0.0) taking (0.0) it (0.0) but (0.0)
it (0.0) dries (0.1) her (0.06) out (0.12)

(c) Single Supervised Self-attention

She (0.0) was (0.0) on (0.0) elavil (0.34) before (0.0)
and (0.0) has (0.0) been (0.0) taking (0.0) it (0.0) but (0.0)
it (0.0) dries (0.28) her (0.17) out (0.21)

(d) Dual Supervised Self-attention



(e) Distribution of attention weights

(f) Sentence prediction scores

Figure 4: Single v.s. dual attention distribution. The color intensity corresponds to the weight given to each word. Attention weight of each word are given in the parenthesis. Single attention-based models (a) and (c) fail to capture sufficient attention weight on the key semantic areas of the sentence. The dual-attention based model where the two attention distributions are combined, accurate weights are assigned (b) and (d).

et al., 2014). The size of the learned character-level embedding are 100 dimensional vectors. The size of LSTM hidden layers for word-level and char-level LSTM are size 300 and 100 respectively. The hidden combined representation h_t was set to size 200; the attention weight layer e_t was set to size 100. The attention-weighted sentence representations TS_S and SS_S, are 200 dimensional vectors and therefore their combination context vector C_S is 400 dimensional. The Entity Prediction Embedding (EPE) L_S is of size k entities that are in BIO format. Hence EPE is a size 19 dimensional binary vector (eighteen entities plus the no entity tag). The final concatenated sentence-level S vector is thus size 419. To avoid over-fitting, we apply a dropout strategy (Ma and Hovy, 2016; Srivastava et al., 2014) of 0.5 for our model. All models were trained with a learning rate of 0.001 using Adam (Kingma and Ba, 2014).

4.3 Results

4.3.1 ADE Assertive Sentence Classification

Table 1 compares our model against two baselines of individual ADE sentence classification models. (i) Similar to (Dernoncourt et al., 2017), LAST is a Bi-LSTM based sentence classification model that uses the last hidden states for sentence composition; (ii) Similar to (Yang et al., 2016), ATTN is a B-LSTM model that used simple attention weights for sentence composition. Our full model, MGADE succeeds to improve the F1 scores by 13.6% over the LAST baseline in testing. We also compare with a model similar to (Zhang et al., 2018) joint-task model based on self-attention. MGADE outperforms their model by 23.0% for sentence classification.

Table 1: ADE sentence classification: F1 scores.

Model	F1
Baseline Individual Models	
LAST (Dernoncourt et al., 2017)	0.66
ATTN (Yang et al., 2016)	0.63
Baseline Joint Model	
(Zhang et al., 2018)	0.61
MGADE	0.75

Table 2: ADE entity recognition: F1 scores.

Model	F1
Baseline Individual Models	
Bi-LSTM (Wunnava et al., 2019)	0.56
Bi-LSTM + CRF (Wunnava et al., 2019)	0.63
Baseline Joint Model	
(Zhang et al., 2018)	0.51
MGADE	0.63

4.3.2 ADE Named Entity Recognition

Table 2 compares our model against the best performing models on MADE1.0 benchmark in the literature (Wunnava et al., 2019) for ADE entity recognition. The entity recognition component of our MGADE is similar to their Bi-LSTM model. MGADE improves the F1 score by 12.5% over their Bi-LSTM only model. Our model achieved comparable results with their Bi-LSTM + CRF combination model. The models with CRF layer predict the label sequence jointly instead of predicting each label individually which is helpful to predict sequences where the label for each word in a sequence depends on the label of the previous

Table 3: Effect of dual-attention layer. † denotes models with single-attention with Task-specific attention removed from Supervised Self-attention model, and vice versa.

	ADE Entity Recognition			**ADE Sentence Classification**		
Model	P	R	F1	P	R	F1
MGADE-SelfA †	0.58	0.52	0.55	0.84	0.55	0.67
MGADE-TaskA †	0.62	0.50	0.55	0.82	0.64	0.72
MGADE-DualA	0.68	0.55	0.61	0.87	0.65	0.74
MGADE	0.70	0.57	0.63	0.86	0.67	0.75

word. Adding an CRF component to our model might further improve the performance of the entity recognition task. We also compare with a model similar to (Zhang et al., 2018) joint-task model based on self-attention. MGADE outperforms their model by 23.5% for entity recognition.

4.3.3 Ablation Analysis

To evaluate the effect of each part in our model, we remove core sub-components and quantify the performance drop in F1 score.

Types of Attention. Table 3 studies the two types of attention we generate: Supervised self-attention (β) and Task-specific attention (α) for composing sentence-level representations. † denotes the models with single-attention. As shown in the table, models that used only a single attention component, be it Supervised Self-Attention based (SS_S) or Task-specific attention based sentence representation (TS_S) achieved the same F1-score for the entity recognition task. However, their sentence classification task performance varies, demonstrating that the two attentions capture different aspects of information in the sentence. The type of attention captured plays a critical role in composing an informative sentence representation. Both single-attention models performed better than the baseline individual sentence-classification models LAST and ATTN (see Table 1). TS_S achieved superior sentence classification performance over SS_S. Intuitively, stronger focus should be placed on the words indicative of the sentence type, and TS_S which emphasizes more on the parts relevant to the ADE sentence classification task is more accurate in identifying ADE sentences.

Single Attention v.s. Dual-Attention. Table 3 studies impact of dual-attention component. As seen, the model with dual-attention sentence representation which combines two attention-weighted sentence representations C_S outperforms the models with single-attention (denoted by †) in both entity recognition and sentence classification tasks.

Label-Awareness. Table 3 studies the effect of adding the label-awareness component in improving the sentence representation. Our full model MGADE, with both dual-attention and label-aware components further improves the performance of sentence classification and entity recognition tasks by 1.0% and 2.0% respectively compared to MGADE-DualA, the model with only dual-attention component.

Case Study. Dual-attention is not only effective in capturing multiple aspects of semantic information in the sentence, but also in reducing the risk of capturing incorrect or insufficient attention when only one of the single attentions (either task-specific or supervised self-attention) is used. Fig 4 shows such an example where single attention, either task-specific or supervised self-attention, fails to capture sufficient attention weight on the key semantic areas of the sentence necessary to make a correct prediction on the sentence. The incorrect distribution of attention weights assigned in the single task-specific and single supervised self-attention (Figures 4a and 4c) is addressed by the dual-attention mechanism. The later corrects the distribution and assigns appropriate weights to the relevant semantic words as in Figures 4b and 4d. In Figures 4e and 4f, we demonstrate the effectiveness of the dual-attention mechanism by plotting attention weight distributions and the sentence prediction scores when specific type of attention is composed into the sentence representation. The bar chart depicts the ADE sentence-level classification confidence scores w.r.t single-attention and dual-attention models and confirms the utility of dual-attention.

5 Conclusion

We propose a dual-attention network for multi-grained ADE detection to jointly identify ADE entities and ADE assertive sentences from medical narratives. Our model effectively supports knowledge sharing between the two levels of granularity, i.e., words and sentences, improving the overall quality of prediction on both tasks. Our solution features significant performance improvements over state-of-the-art models on both tasks. Our MGADE architecture is pluggable, in that other sequential learning models including BERT (Devlin et al., 2019) or other models for sequence labelling and text classification could be substituted in place of the Bi-LSTM sequential representation learning model. We leave this enhancement of our model and its study to future work.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. Neural networks for joint sentence classification in medical paper abstracts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 694–700. Association for Computational Linguistics.

Shantanu Dev, Shinan Zhang, Joseph Voyles, and Anand S Rao. 2017. Automated classification of adverse events in pharmacovigilance. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 905–909. IEEE.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Molla S. Donaldson, Janet M. Corrigan, Linda T. Kohn, and Editors. 2000. *To err is human: building a safer health system*, volume 6. National Academies Press.

Harsha Gurulingappa, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2011. Identification of adverse drug event assertive sentences in medical case reports. In *First international workshop on knowledge discovery and health care management (KD-HCM), European conference on machine learning and principles and practice of knowledge discovery in databases (ECML PKDD)*, pages 16–27.

Rave Harpaz, Alison Callahan, Suzanne Tamang, Yen Low, David Odgers, Sam Finlayson, Kenneth Jung, Paea LePendu, and Nigam H Shah. 2014. Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug safety*, 37(10):777–790.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–1780.

Trung Huynh, Yulan He, Alistair Willis, and Stefan Rüger. 2016. Adverse drug reaction classification with deep neural networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 877–887.

Abhyuday Jagannatha, Feifan Liu, Weisong Liu, and Hong Yu. 2019. Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (made 1.0). *Drug safety*, 42(1):99–111.

Abhyuday N Jagannatha and Hong Yu. 2016a. Bidirectional rnn for medical event detection in electronic health records. In *Proceedings of the conference. ACL. North American Chapter. Meeting*, volume 2016, page 473. NIH Public Access.

Abhyuday N. Jagannatha and Hong Yu. 2016b. Structured prediction models for rnn based sequence labeling in clinical text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 856. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Feifan Liu, Abhyuday Jagannatha, and Hong Yu. 2019. Towards drug safety surveillance and pharmacovigilance: current progress in detecting medication and adverse drug events from electronic health records.

Lemao Liu, Masao Utiyama, Andrew M. Finch, and Eiichiro Sumita. 2016. Neural machine translation with supervised attention. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pages 3093–3102. ACL.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. 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*. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers.

Chandra Pandey, Zina M. Ibrahim, Honghan Wu, Ehtesham Iqbal, and Richard J. B. Dobson. 2017. Improving RNN with attention and embedding for adverse drug reactions. In *Proceedings of the 2017 International Conference on Digital Health, London, United Kingdom, July 2-5, 2017*, pages 67–71.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP).

Marek Rei and Anders Søgaard. 2019. Jointly learning to label sentences and tokens. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6916–6923.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15:1929–1958.

Ahmad P Tafti, Jonathan Badger, Eric LaRose, Ehsan Shirzadi, Andrea Mahnke, John Mayer, Zhan Ye, David Page, and Peggy Peissig. 2017. Adverse drug event discovery using biomedical literature: a big data neural network adventure. *JMIR medical informatics*, 5(4):e51.

Peng Wang, Lingqiao Liu, Chunhua Shen, Zi Huang, Anton van den Hengel, and Heng Tao Shen. 2017. Multi-attention network for one shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2721–2729.

Susmitha Wunnava, Xiao Qin, Tabassum Kakar, Cansu Sen, Elke A Rundensteiner, and Xiangnan Kong. 2019. Adverse drug event detection from electronic health records using hierarchical recurrent neural networks with dual-level embedding. *Drug safety*, 42(1):113–122.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Shinan Zhang, Shantanu Dev, Joseph Voyles, and Anand S Rao. 2018. Attention-based multi-task learning in pharmacovigilance. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2324–22328. IEEE.