Improving Distantly Supervised Relation Extraction using Word and Entity Based Attention

Sharmistha Jat[1] *, **Siddhesh Khandelwal**[2] *†, **Partha Talukdar**[1]
[1]Indian Institute of Science, Bangalore
[2]University of British Columbia
{sharmisthaj,ppt}@iisc.ac.in,skhandel@cs.ubc.ca

1 Introduction

Classifying the semantic relationship between two entities in a sentence is termed as Relation Extraction (RE). RE from entity mentions is an important step in various Natural Language Processing tasks, such as, knowledge base construction, question-answering etc. Supervised methods have been successful on the relation extraction task [2, 18]. However, the extensive training data necessary for supervised learning is expensive to obtain and therefore restrictive in a Web-scale relation extraction task. To overcome this challenge, [6] proposed a Distant Supervision (DS) method for relation extraction to help automatically generate new training data by taking an intersection between a text corpora and knowledge base. However, the DS assumption is too strong, and may introduce noise such as false negative samples due to missing facts in knowledge base. In order to address this challenge, DS has been modeled as Multi-Instance Multi-Label (MIML) problem [14]. More recently, neural models for DS have been proposed [17, 12]. In this paper, we define 'instance' as a sentence containing an entity-pair, and 'instance set' as a set of sentences containing the same entity-pair.

It was observed by [17] that 50% of the sentences in the Riedel2010 DS dataset [9] had 40 or more words in them. We observe that not all the words in these long sentences contribute towards expressing the given relation. In this work, we formulate various word attention mechanisms to help the relation extraction model focus on the right context in a given sentence.

The MIML assumption states that in an instance set corresponding to an entity pair, at least one sentence in that set should express the true relation assigned to the set. However, we observe that this is not always true in currently available benchmark datasets for RE in the distantly supervised setting. In particular, current datasets have noise in the *test* set, which impedes the right comparison of models. To address this challenge, we build the Google-IISc Distant Supervision (GIDS) dataset, a new dataset for distantly-supervised relation extraction. GIDS is seeded from the Google relation extraction corpus [11]. This new dataset addresses an important shortcoming in distant supervision evaluation, and makes automatic evaluation in this setting more reliable.

In summary, our contributions are: (a) we introduce the Google-IISc Distant Supervision (GIDS) dataset, a new dataset for distantly-supervised relation extraction; (b) we propose two novel word attention based models for distant supervision, viz., BGWA, a BiGRU-based word attention model, and EA, an entity-centric attention model; and (c) we show efficacy of combining new and existing relation extraction models using a weighted ensemble model.

Our code and datasets are publicly available at https://github.com/SharmisthaJat/RE-DS-Word-Attention-Models.

2 Proposed Methods

In this section, we describe propose models: (1) BGWA(Section 2.1); (2) EA(Section 2.2); and (3) Weighted voting ensemble model (Section 2.3).

2.1 Bi-GRU based Word Attention Model (BGWA)

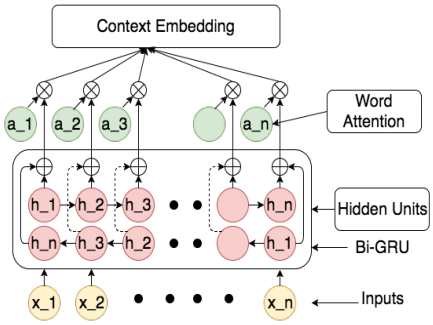
Consider the sentence, *Former President Barack **Obama** was born in the city of **Honolulu**, capital of the U.S. state of Hawaii*, expressing *bornIn(Person, City)* relation between the entity pair *(Obama, Honolulu)*. In the sentence, the phrase *"was born in"* is helps in identifying the correct relation in the sentence. It is conceivable that identifying such key phrases or words will be helpful in improving relation extraction performance. BGWA uses an attention mechanism over words to identify such key

*Equal Contribution to this work

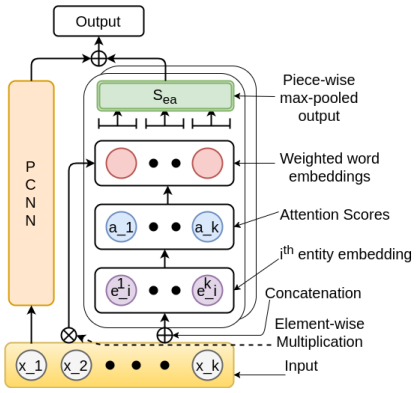
†This research was conducted during the author's research assistantship at Indian Institute of Science

6th Workshop on Automated Knowledge Base Construction (AKBC) at NIPS 2017, Long Beach, CA, USA.

Figure 1: Proposed word attention models



(a) Bi-GRU word attention (BGWA) model, (Section 2.1)



(b) Entity Attention (EA) Model (Section 2.2)

phrases. To the best of our knowledge, there has been no prior work on using word attention in the distant supervision setting.

To identify such key words, BGWA leverages the ability of Bidirectional Gated Recurrent Unit (Bi-GRU). GRUs [3] capture long range structural dependencies, and thereby give better context embedding for a sentence. Additionally, Bi-GRU provides richer representations, by capturing dependencies in both directions as it runs a GRU in both the forward and reverse direction of the word ordering.

Architecture of BGWA is shown in Figure 1a. Assume a sentence consists of k words. Each word is represented using an embedding in a d-dimensional input space, i.e., $x_i \in \mathbb{R}^{d \times 1}$. Each sentence is processed using two GRUs [3], one processing it in the forward direction and the other in the reverse direction. Representation $w_i \in \mathbb{R}^{1 \times g}$ of the i^{th} word is obtained by concatenating individual hidden representations of length g / 2, provided by the forward (h_i^f) and backward (h_i^b) GRUs. We define u_i, the degree of relevance of the i^{th} word, as follows.

$$w_i = [h_i^f, h_i^b]; u_i = w_i \times A \times r; a_i = \frac{\exp(u_i)}{\sum_{j=1}^{k} \exp(u_j)}; \hat{w}_i = a_i \times x_i$$

where $A \in R^{g \times g}$ is a square matrix and $r \in \mathbb{R}^{g \times 1}$ is a relation vector. Both A and r are learned. Attention value a_i is calculated by taking softmax over $\{u_i\}$ values. The attention values a_i are used to generate a weighted representation for each word $\hat{w}_i \in \mathbb{R}^{1 \times g}$.

Despite the widespread use of weighted sum to obtain context embeddings in attention based settings, we choose the piecewise max pooling method to retain structural properties of context before, between, and after the entity pair [17]. Let $I_1 = [1, p_1 - 1]$, $I_2 = [p_1, p_2]$, and $I_3 = [p_2 + 1, k]$ be the indices of words occurring to the left of the first entity, words occurring between and including the two entities, and words occurring to the right of the second entity, respectively. Embedding $s_j \in \mathbb{R}^{|I_j| \times g}$ for words in segment I_j is obtained by stacking attention-weighted embeddings for all words in that segment. We define s_1, s_2, and s_3 as follows.

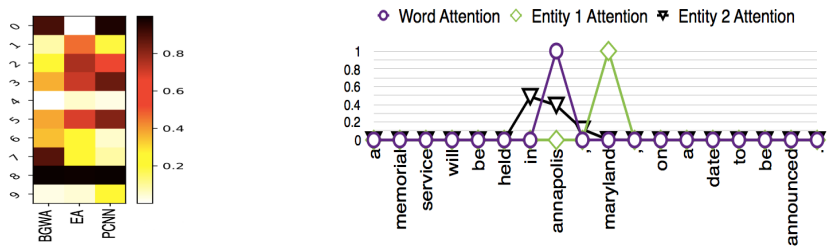
$$s_1 = [\hat{w}_1; \cdots; \hat{w}_{p_1 - 1}], s_2 = [\hat{w}_{p_1}; \cdots; \hat{w}_{p_2}], s_3 = [\hat{w}_{p_2 + 1}; \cdots; \hat{w}_k]$$

We then apply a max pooling over each of the three segments created by the two entities. Final embedding $s_{wa} \in \mathbb{R}^{1 \times 3g}$ of the sentence is then be processed through a linear layer to yield probabilities for each relation.

$$s_{wa} = \langle \max(s_1), \max(s_2), \max(s_3) \rangle$$

2.2 Entity Attention (EA) Model

Let us once again consider the example sentence from Section 2.1 involving entity pair *(Obama, Honolulu)*. In the sentence, for entity *Obama*, the word *President* helps in identifying that the entity



(a) Model Confidence Scores *(b) Word Attention*

Figure 2: (a) Confidence scores (indicated by color intensity, darker is better) of models on true labels of 10 randomly sampled instance sets from Google-IISc Distant Supervision. Rows represent the instance sets and columns represent the model used for prediction. The heatmap shows complementarity of these models in selecting the right relation. Motivated by this evidence, the proposed Ensemble method trains to learn weights and combines the three models, viz., Word Attention (BGWA), Entity Attention (EA) and PCNN. (b) BGWA (Word attention) and Entity Attention (EA) values for an example sentence between entity pair maryland and annapolis and relation location_in. X-axis shows the sentence words and y-axis shows the attention scores. Please see Section 4 for more discussion.

is a person. This extra information helps in narrowing down the relation possibilities by looking only at the relations that occur between a person and a city. [12] proposed an entity attention model for *supervised* relation extraction with a single sentence as input to the model. We modify and adapt that model for the *distant supervision* setting and propose Entity Attention (EA) which works with a bag of sentences. For a given bag of sentences, the learning is done using the setting proposed by [17], wherein the sentence with highest probability of expressing a relation in a bag is selected to train the model in each iteration. Architecture of EA is shown in Figure 1b.

The EA model has two components: 1) PCNN layer and 2) Entity Attention Layer. Note that although the authors [12] used a CNN followed by Max Pooling in their work, we instead use the PCNN model [17] due to its effectiveness in a distantly supervised setting. Consider a sentence consisting of k words $\langle x_1, x_2 \dots x_k \rangle$, where each $x_i \in \mathbb{R}^{1\times d}$ is a word embedding and $\{e_1, e_2\}$, $e_j \in \mathbb{R}^{1\times d}$ are the embeddings for the two entities. The PCNN layer is applied on the words in the sentence (x_i's) [17]. The entity-specific attention for each word is calculated as follows.

$$u_{i,j} = [x_i, e_j] \times A_j \times r_j, i \in [1, k], j \in \{1, 2\}$$

Here, A_j and r_j are learned parameters. The $u_{i,j}$ are normalized using a softmax function to generate $a_{i,j}$, the attention scores. Although [12] used a weighted sum approach over the word embeddings to obtain the sentence context vector, we instead use piecewise max pooling [17] over the weighted word embeddings due to its effectiveness in a distantly supervised setting (As explained in Section 2.1). Let the two entities be located at positions p_1 and p_2 in the sentence, where $1 \le p_1 \le p_2 \le k$. Where $s_j, j \in \{1, 2\}$ is calculated as follows,

$$s_j = \langle \max(\{a_{0,j} * x_0, ..., a_{p_1-1,j} * x_{p_1-1}\}), \max(\{a_{p_1,j} * x_{p_1}, ..., a_{p_2,j} * x_{p_2}\}), \max(\{a_{p_2+1,j} * x_{p_2+1}, ..., a_{p_k,j} * x_{p_k}\})$$

The sentence context vector can be represented as.

$$s_{ea} = \langle \max(s_1), \max(s_2), \max(s_3) \rangle$$

The output from the PCNN layer and the entity attention layer are concatenated and then passed through a linear layer to obtain probabilities for each relation.

2.3 Bring it all together: Ensemble Model

We note that the two models discussed in previous sections, BGWA, EA and PCNN have complementary strengths. PCNN extracts high-level semantic features from sentences using CNN. Most effective features are then selected using a piecewise max-pooling layer. Entity-based attention (Section 2.2) helps in highlighting important relation words with respect to each of the entities present in the sentence, thus complimenting the PCNN-based features. Going beyond the entity-centric words, we observe that not all words in a sentence are equally important from relation extraction perspective. The BGWA model (Section 2.1) addresses this aspect by selecting words relevant to a relation in a sentence.

Table 1: Examples of Noise in dataset. Sample 1,2 are incorrectly labelled with NA relation in the test set due to missing facts in Knowledge Base (KB). While, Sample 3's single sentence in the instance set does not support the KB relation.

Entity 1	Entity 2	Test Set Label	Classified Relation
1. Marlborough	New Hampshire	NA	/location/location/contains
2. Katie Couric	CBS	NA	/business/person/company

Entity 1	Entity 2	Test Set Label	Instance Set
3. Gary Sheffield	Florida	/people/person/ place_lived	others who have already indicated they will wear no. 42 include ken griffey jr. of cincinnati,florida's dontrelle willis, carlos lee of houston, derrek lee of the cubs and detroit's gary_sheffield .

Table 2: Dataset Statistics

(a) Statistics of the new GIDS dataset. Please see Section 3 for more details.

Relation - Class	No. sentences	No. entity-pair
perGraduatedInstitution	4456	2624
perHasDegree	2969	1434
perPlaceOfBirth	3356	2159
perPlaceOfDeath	3469	1948
NA	4574	2667

(b) Statistics of various datasets used in the paper.

Dataset	# relation	# sentences	# entity-pair
Reidel2010 Dataset with development set			
Train	53	455,771	233,064
Dev	53	114,317	58,635
Test	53	172,448	96,678
GIDS Dataset			
Train	5	11297	6498
Dev	5	1864	1082
Test	5	5663	3247

In Figure 2a, we plot the confidence scores of various models on the true labels of 10 randomly selected instance sets from Google-IISc Distant Supervision dataset. From this figure, we observe that the proposed methods are able to leverage signals from the entity and word attention models, even when the PCNN model is incorrect (light colored cell in the last column). This validates our assumption and motivates ensemble approach to efficiently combine these complementary models.

We combined the predictions of all the three models using a weighted voting ensemble. The weights of this model were learned using linear regression. More complicated regression methods did not improve the results greatly. We also experimented with a jointly learned neural ensemble by concatenating the features of all models after pooling layer followed by a linear layer. In our experiments weighted voting ensemble method gave better results than jointly learned model.

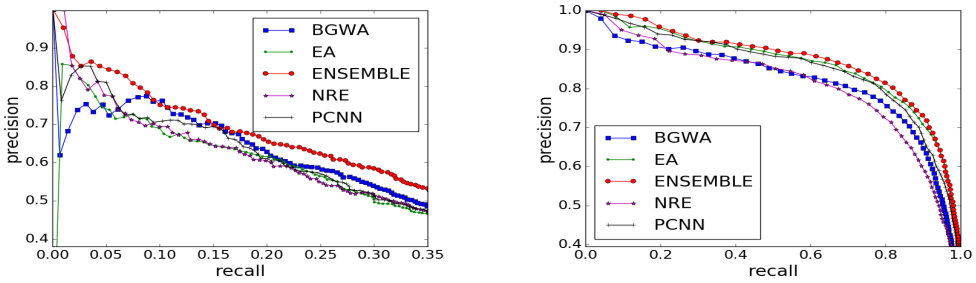
3 GIDS: A New Dataset for Relation Extraction using Distant Supervision

Several benchmark datasets for relation extraction using distant supervision (DS) exist [9, 6]. DS is used to create both train and test sets in all of these datasets, resulting in introduction of noise in the dataset. While training noise in distant supervision is expected, noise in the test data is troublesome as it may lead to incorrect evaluations.

There are two kinds of noise added due to distant supervision assumption, (a) samples with incorrect labels due to missing KB fact, (b) samples with no instance supporting the KB fact, some examples are listed in Table 1. Previous benchmark datasets in this area suffer from these drawbacks. In order to overcome these challenges, we develop newdataset (GIDS), a new dataset for relation extraction using distant supervision. Statistics of the new dataset are summarized in Table 2a. To alleviate noise in DS setting, we make sure that labelled relation is correct and for each instance set in GIDS, there is at least one sentence in that set which expresses the relation assigned to that set. We start with the human-judged Google Relation Extraction (RE) corpus [11]. This corpus consists of 5 binary relations: (1) *perGraduatedFromInstitution*, (2) *perHasDegree*, (3) *perPlaceOfBirth*, (4) *perPlaceOfDeath*, and (5) none of the above (NA).

We constructed the GIDS dataset using the following process. Let $D_{\text{GRE}} = \{(\mathbf{x}_i, e_{i1}, e_{i2}, r_i)\}$ be the Google RE corpus, where the i^{th} sentence $\mathbf{x}_i$ is annotated as expressing relation r_i between the two entities e_{i1} and e_{i2} in the sentence. r_i is one of the five relations mentioned above. Now, for each $(\mathbf{x}_i, e_{i1}, e_{i2}, r_i) \in D_{\text{GRE}}$, we perform the following:

- Perform web search to retrieve documents containing the two entities e_{i1} and e_{i2}. From retrieved documents, select multiple text snippets containing the entities. Each snippet is restricted to contain at most 500 words. Let $S_i = \{(\mathbf{s}_i)\}$ be the set of such snippets.
- Let $S_i^{'} = \{\{\mathbf{x}_i\} \cup S_i\}$. We create a new instance set $B_i = \{(S_i^{'}, e_{i1}, e_{i2}, r_i)\}$ for distance supervision which consists of the set of instances (sentences or snippets) $S_i^{'}$, where the entities e_{i1} and e_{i2} are mentioned in each instance. Label r_i is applied over the entire set B_i.



(a) Riedel2010 dataset with dev partition *(b) GIDS dataset*

Figure 3: Precision-recall curves of various models over two datasets: (a) for the 53 relation classification in the Riedel2010 dataset with development split. We partitioned Reidel2010's train set into a new train (80%) and development set (20%); and (b) for the 5 relation classification in the GIDS dataset. Please see Section 4 for more details.

$D_{\text{GIDS}} = \{B_i\}$ is the new GIDS dataset. Here, each set B_i is guaranteed to contain at least one sentence ($\mathbf{x}_i$) which expresses the relation r_i assigned to that set. We note that such guarantee was not available in previous DS benchmarks. We divided this dataset into train (60%), development (10%) and test (30%) sets, such that there is no overlap among entity-pairs of these sets. Unlike currently available datasets, the availability of development dataset helps in performing model selection in a principled manner for relation extraction. In [9] and subsequent work, a manual evaluation was done by validating the top 1000 confident predictions. This manual evaluation was necessary due to the noise in the test data. GIDS gets past such cumbersome manual evaluation and makes reliable automated evaluation in distantly-supervised relation extraction a reality.

4 Experiments and Results

We validate the effectiveness of the proposed word attention, entity attention and ensemble models on multiple datasets. Table 2b summarizes information about the datasets. **Riedel2010** was created by aligning Freebase relations with the NYT corpus [9, 4, 14, 5] [3]. Construction of the new **GIDS** dataset is described in Section 3.

Evaluation Metrics: Following [5], we use held-out evaluation scheme. The performance of each model is evaluated using Precision-Recall (PR) curve. Following code implementation by [5], the PR curve is calculated by first sorting the predicted relations based on the confidence of prediction, followed by calculating the precision and recall on an ever expanding set by including each prediction from the highest confidence (low recall) to the lowest confidence (full recall). We compare with the following baseline methods:

- **PCNN** [17]: Piecewise Convolution Neural Network (PCNN) for relation extraction is an effective model for relation extraction (more details in [17]). For GIDS dataset we used development dataset to select learning rate of 2.0 and set maximum number of epochs as 40.
- **NRE** [5]: This is the current state-of-the-art method on the Riedel2010 dataset. It uses sentence attention to select relevant sentences from the bag of sentences for relation prediction for a given entity-pair. Results for the best performing model (PCNN+ATT) on the Riedel2010 dataset were taken from author provided code [5]. For GIDS dataset, we perform model selection using development data by varying learning rate (selected value: 1.4) and number of iterations (selected value: 30).

Model Parameters: We use $d_w = 50$ dimensional word embeddings, initialized using the Word2Vec vectors from [5] dataset. Position feature embeddings of length $d_p = 5$ are randomly initialized and learned while training our models. The parameters used for the various models are mentioned below.

- *PCNN*: Following [17], we implemented this baseline with 230 filters in a single convolution layer with a window size of 3, followed by piecewise max-pooling.
- *Entity Attention Network (EA)* (Section 2.1): The attention is applied on word embeddings after appending entity embeddings to them. Therefore, the size of the final word embedding is $2\times(d_w+2\times d_p) = 120$. Piecewise pooling is applied to the weighted word representations to obtain a sentence representation of size 360 for each pipeline.

[3] Dataset downloaded from `https://github.com/thunlp/NRE`

- *Bidirectional-GRU Word Attention (BGWA)* (Section 2.1): The outputs Bi-GRUs (460 features) is the input to the word attention module. Piecewise pooling is performed on the weighted GRU embeddings obtained from the attention module to obtain a sentence representation of size 1380.

The PCNN baseline outperforms traditional baselines like MIML-RE and hence we use PCNN as a starting baseline to compare with proposed models. We use SGD algorithm to learn models with learning rate set to 0.1 (unless otherwise stated) on a batch size of 50. Dropout [13] of 0.5 is applied before the linear layer during training. All the models implemented by us were developed in PyTorch[4]. The experiments were run on GeForce GTX 1080 Ti using NVIDIA-CUDA. The ensemble model weights were learnt using linear regression [8].

Results: Figure 3 shows the precision-recall curve for baseline and proposed algorithms on two datasets, Riedel2010 and GIDS. From the plots, we observe that the proposed BGWA model outperforms PCNN and NRE in the Riedel2010 dataset, while EA outperforms PCNN and NRE in the GIDS dataset. We also note that EA is comparable to NRE in the Riedel2010 dataset, while BGWA is comparable to NRE in the GIDS dataset. Ensemble (Section 2.3), a combination of the proposed models – BGWA and EA – along with PCNN significantly outperforms all models on all datasets. This indicates that combined clues about the entity-type and context from the model combination can result in better prediction. Our attention-based models help Ensemble focus on relevant words in the sentences. In Figure 3a, we plot NRE baseline result obtained from the published paper. This model was trained on full train, which gives this baseline an advantage over other methods which use 80% train data for learning.

We visualize the attention values of our models in Figure 2b. It can be observed that the entity attention rightly chooses words like '*in*', '*,*' and the entity names. As the word attention is applied on GRU hidden layer output, a high attention value for the hidden layers after processing the word '*annapolis*' indicates that the sentence has rich context around the first entity to indicate *location_in* relation. In conclusion, the word attention models rightly choose the useful words in context and help in improving relation extraction results.

5 Related Work

A large proportion of the work in this field has aimed to relax the strong assumptions that the original DS model made. [9] introduced the *expressed-at-least-once* assumption in a factor graph model as an aggregating mechanism over mention level predictions. Work by [4, 14, 10] are crucial increments to [9] to improve RE further. Deep learning models proposed by [18, 17] have reduced dependence on manual feature extraction. [17] proposed a Piecewise Convolutional Neural Network (PCNN) model to tackle the issue of hand-crafted feature engineering. [19] aimed to leverage *inter-sentence* information for relation extraction in a ranking model. [5] uses sentence attention to select relevant sentences from the instance-set. Recently, work by [16] exploit the connections between relation (class ties) to improve relation extraction performance. [7] use inter-instance-set couplings for relation extraction in multi-task setup to improve the performance. Attention models learn the importance of a feature in the supervised task through back-propogation. Attention mechanisms in neural networks have been successfully applied to a variety of problems, like machine translation [1], image captioning [15], supervised relation extraction [12], distantly-supervised relation extraction [19] etc.

In our work, we propose models which are complementary to previously proposed models. We focus on selecting the right words in a sentence using word and entity-based attention mechanism. We further combine the existing approaches with a weighted voting ensemble to improve results.

6 Conclusion

In this paper, we make following contributions: (a) we introduce the GIDS dataset, a new dataset for distantly-supervised relation extraction; (b) we propose two novel word attention based models for distant supervision, viz., BGWA, a BiGRU-based word attention model, and EA, an entity-centric attention model; and (c) we show efficacy of combining new and existing relation extraction models using a weighted ensemble model. GIDS dataset removes test data noise present in all previous distance supervision benchmark datasets, making credible automatic evaluation possible. Combining proposed methods with attention-based sentence selection methods is left as future work. Our code and datasets are publicly available at `https://github.com/SharmisthaJat/RE-DS-Word-Attention-Models`.

[4] `http://pytorch.org/`

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Razvan C Bunescu and Raymond J Mooney. A shortest path dependency kernel for relation extraction. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*, pages 724–731. Association for Computational Linguistics, 2005.

[3] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[4] Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 541–550. Association for Computational Linguistics, 2011.

[5] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2124–2133, Berlin, Germany, August 2016. Association for Computational Linguistics.

[6] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics, 2009.

[7] Tushar Nagarajan, Sharmistha , and Partha Talukdar. Candis: Coupled and attention-driven neural distant supervision. *arXiv preprint arXiv:1710.09942*, 10 2017.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[9] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer, 2010.

[10] Alan Ritter, Luke Zettlemoyer, Oren Etzioni, et al. Modeling missing data in distant supervision for information extraction. *Transactions of the Association for Computational Linguistics*, 1:367–378, 2013.

[11] Shaohua Sun, Ni Lao, Rahul Gupta and Dave Orr. 50,000 Lessons on How to Read: a Relation Extraction Corpus. `https://research.googleblog.com/2013/04/50000-lessons-on-how-to-read-relation.html`, 2013. [Online; accessed 15-NOV-2017].

[12] Yatian Shen and Xuanjing Huang. Attention based convolutional neural network for semantic relation extraction. In *COLING*, 2016.

[13] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[14] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. Multi-instance multi-label learning for relation extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465. Association for Computational Linguistics, 2012.

[15] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. pages 2048–2057, 2015.

[16] Hai Ye, Wenhan Chao, and Zhunchen Luo. Jointly extracting relations with class ties via effective deep ranking. *CoRR*, abs/1612.07602, 2016.

[17] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *EMNLP*, 2015.

[18] Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344, 2014.

[19] Hao Zheng, Zhoujun Li, Senzhang Wang, Zhao Yan, and Jianshe Zhou. Aggregating inter-sentence information to enhance relation extraction. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.