Neural Attention Model for Classification of Sentences that Support *Promoting/Suppressing* Relationship

Yuta Koreeda, Toshihiko Yanase, Kohsuke Yanai, Misa Sato, Yoshiki Niwa
Research & Development Group, Hitachi, Ltd.
{yuta.koreeda.pb, toshihiko.yanase.gm, kohsuke.yanai.cs,
misa.sato.mw, yoshiki.niwa.tx}@hitachi.com

Abstract

Evidences that support a claim "a subject phrase *promotes* or *suppresses* a value" help in making a rational decision. We aim to construct a model that can classify if a particular evidence supports a claim of a *promoting/suppressing* relationship given an arbitrary subject-value pair. In this paper, we propose a recurrent neural network (RNN) with an attention model to classify such evidences. We incorporated a word embedding technique in an attention model such that our method generalizes for never-encountered subjects and value phrases. Benchmarks showed that the method outperforms conventional methods in evidence classification tasks.

1 Introduction

With recent trend of big data and electronic records, it is getting increasingly important to collect evidences that support a claim, which usually comes along with a decision, for rational decision making. Argument mining can be utilized for this purpose because an argument itself is an opinion of the author that supports the claim, and an argument usually consists of evidences that support the claim. Identification of a claim has been rigorously studied in argument mining including extraction of arguments (Levy et al., 2014; Boltui and najder, 2014; Sardianos et al., 2015; Nguyen and Litman, 2015) and classification of claims (Sobhani et al., 2015).

Our goal is to achieve classification of positive and negative effects of a subject in a form "a subject phrase $\mathcal{S}$ *promotes/suppresses* a value $\mathcal{V}$." For example, given a subject $\mathcal{S}$ = `gambling`, a value $\mathcal{V}$ = `crime` and a text $\mathcal{X}$ = `casino increases theft`, we can say that $\mathcal{X}$ supports a claim of `gambling` ($\mathcal{S}$) `promotes crime` ($\mathcal{V}$) relationship. Such a technique is important because it allows extracting both sides of an opinion to be used in decision makings (Sato et al., 2015).

We take a deep learning approach for this evidence classification, which has started to outperform conventional methods in many linguistic tasks (Collobert et al., 2011; Shen et al., 2014; Luong et al., 2015). Our work is based on a neural attention model, which had promising result in a translation task (Bahdanau et al., 2015) and in a sentiment classification task (Zichao et al., 2016). The neural attention model achieved these by focusing on important phrases; e.g. when $\mathcal{V}$ is `economy` and $\mathcal{X}$ is `Gambling boosts the city's revenue.`, the attention layer focuses near the phrase `boosts the city's revenue`.

The neural attention model was previously applied to aspect-based sentiment analysis (ABSA) (Yanase et al., 2016), which has some similarity to the evidence classification in that it classifies sentimental polarities towards a subject $\mathcal{S}$ given an aspect (corresponding to $\mathcal{V}$) (Pontiki et al., 2015). A limitation of (Yanase et al., 2016) was that the learned attention layer is tightly attached to each $\mathcal{S}$ or $\mathcal{V}$ and does not generalize for never-encountered subjects/values. This means that it requires manually labeled data for all possible subjects and values, which is not practicable. Instead, when we train a model to classify an evidence that supports a claim of a relationship between, for example, `gambling` and `crime`, we want the same learned model to work for other $\mathcal{S}$ and $\mathcal{V}$ pairs such as `smoking` and `health`. In other words, we want the model to learn *how* to classify evidences that support a relationship of $\mathcal{S}$ and $\mathcal{V}$, rather than learning the relationship itself.

In this paper, we propose a neural attention

Proceedings of the 3rd Workshop on Argument Mining, pages 76–81,
Berlin, Germany, August 7-12, 2016.

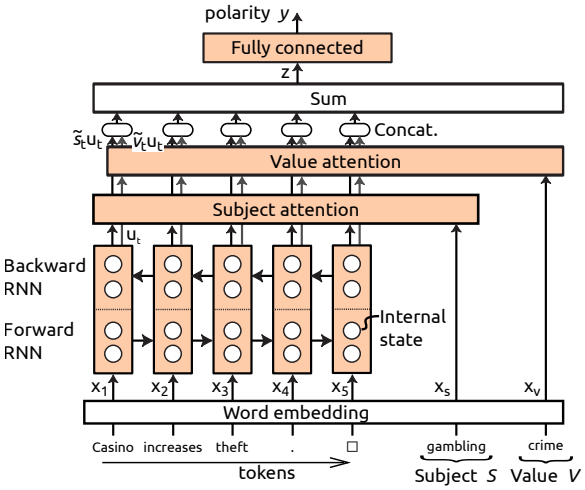


Figure 1: Structure of the proposed bi-directional RNN with word embedding-based attention layer. Colored units are updated during training.

Subject $\mathcal{S}$	Value $\mathcal{V}$ (# of *promoting* / *suppressing* / total labels)
Training data	
national lottery	economy (88 / 57 / 145), regressive tax (4 / 1 / 5)
sale of human organ	moral (0 / 6 / 6)
generic drug	cost (32 / 87 / 119), poverty (0 / 1 / 1)
cannabis	economy (61 / 7 / 68), medicine (215 / 68 / 283)
tourism	economy (142 / 11 / 153), corruption (10 / 3 / 13)
Test data	
smoking	income (36 / 33 / 69), disease (158 / 1 / 159)
violent video game	crime (36 / 7 / 43), moral (7 / 14 / 21)

Table 1: Subject phrases and value phrases in the dataset

model that can learn to focus on important phrases from text even when $\mathcal{S}$ and $\mathcal{V}$ are never encountered, allowing the neural attention model to be applied to the evidence classification. We extend the neural attention model by modeling the attention layer using a distributed representation of words in which similar words are treated in a similar manner. We also report benchmarks of the method against previous works in both neural and lexicon-based approaches. We show that the method can effectively generalize to an evidence classification task with never-encountered phrases.

2 Neural Attention Model

Given a subject phrase $\mathcal{S}$, a value phrase $\mathcal{V}$, and a text $\mathcal{X}$, our model aims to classify whether $\mathcal{X}$ supports $\mathcal{S}$ *promotes* or *suppresses* $\mathcal{V}$. A text $\mathcal{X}$ is a sequence of word tokens, and the classification result is outputted as a real value $y \in [0.0, 1.0]$ that denotes the *promoting*/*suppressing* polarity; i.e., $\mathcal{X}$ has a higher chance of supporting the *promoting* claim if it is nearer to 1.0 and the *suppressing* claim if it is nearer to 0.0.

Our method is shown in Figure 1. First of all, we apply skip-gram-based word embedding (Mikolov et al., 2013) to each token in $\mathcal{X}$ and obtain a varying-length sequence of distributed representations $X = \mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_T$, where T is the number of tokens in the sentence. This is to allow words with similar meaning to be treated in a similar manner.

We also apply word embedding to $\mathcal{S}$ and $\mathcal{V}$ to obtain $\mathbf{x}_s$ and $\mathbf{x}_v$ respectively. This is a core idea on making attention model generalize to first encountered words. In case there exists more than one word in $\mathcal{S}$ and $\mathcal{V}$, we take an average of word embedding vectors.

Next, the word vector sequence X is inputted to a recurrent neural network (RNN) to encode contextual information into each token. The RNN calculates an output vector for each $\mathbf{x}_t$ at token position t. We use a bi-directional RNN (BiRNN) (Schuster and Paliwal, 1997) to consider both forward context and backward context. A forward RNN processes tokens from head to tail to obtain a forward RNN-encoded vector $\overrightarrow{\mathbf{u}_t}$, and a backward RNN processes tokens from tail to head to obtain a backward RNN-encoded $\overleftarrow{\mathbf{u}_t}$. The output vector is $\mathbf{u}_t = \overrightarrow{\mathbf{u}_t}||\overleftarrow{\mathbf{u}_t}$, where $||$ is the concatenation of vectors. We tested the method with long short-term memory (LSTM) (Sak et al., 2014) and gated recurrent units (GRUs) (Cho et al., 2014) as implementations of RNN units.

Lastly, we filter tokens with $\mathcal{S}$ and $\mathcal{V}$ to determine the importance of each token and to extract information about the interactions of $\mathcal{S}$ and $\mathcal{V}$. In the attention layer, attention weight $s_t \in \mathbb{R}$ at each token t is calculated using subject phrase vector $\mathbf{x}_s$. We model attention with Equation (1) in which W_s is a parameter that is updated alongside the RNN during the training.

$$s_t = \mathbf{x}_s^\top W_s \mathbf{u}_t \tag{1}$$

Then, we take the softmax over all tokens in a sentence for normalization.

$$\tilde{s}_t = \frac{\exp(s_t)}{\sum_j \exp(s_j)} \tag{2}$$

Parameter	BiRNN	BiRNN+ATT
Dropout rate	0.7	0.5
Learning rate	0.00075	0.0017
RNN model	GRU	LSTM
RNN state size	128	64
Mini-batch size	16	32
Training epochs	6	17

Table 2: Hyperparameters of BiRNN and BiRNN+ATT (our method)

	Average AUC-PR	AUC-ROC	Macro prec.	Accuracy
BiRNN+ATT	**0.59**	**0.64**	**0.62**	**0.51**
BiRNN	0.57	0.59	0.54	0.45
BoM	0.58	0.57	0.49	0.22
BoW	0.56	0.61	0.56	0.42

Table 3: Performance of the classifiers. The best result for each metric is shown bold.

The attention $\tilde{v}_t \in \mathbb{R}$ for the value vector is calculated likewise using a parameter W_v. $\tilde{s}_t$ and $\tilde{v}_t$ are used as the weight of each token $\mathbf{u}_t$ to obtain sentence feature vector $\mathbf{z}$.

$$\mathbf{z} = \sum_t (\tilde{s}_t \mathbf{u}_t || \tilde{v}_t \mathbf{u}_t) \quad (3)$$

Finally, the polarity y of the claim is calculated two-layered fully-connected perceptrons with logistic sigmoid functions.

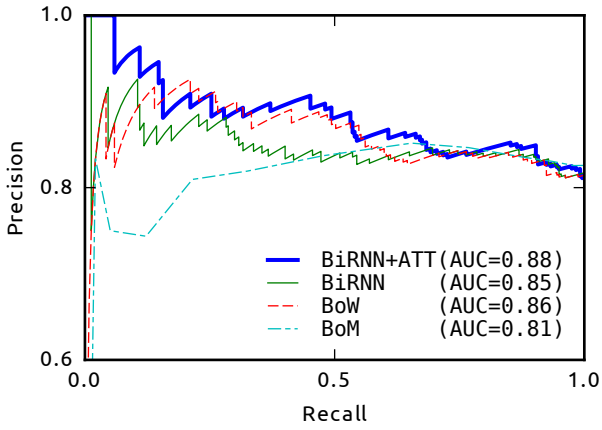
The model is trained by backpropagation using cross entropy as the loss and AdaGrad as the optimizer (Duchi et al., 2011). During training, parameters of fully-connected layers, RNN, W_s, and W_v are updated. Note that $\mathbf{x}_s$, $\mathbf{x}_v$ are not updated unlike (Yanase et al., 2016). Dropout (Srivastava et al., 2014) is applied to the input and output of the RNN and gradient norm is clipped to 5.0 to improve the stability.

3 Experiments

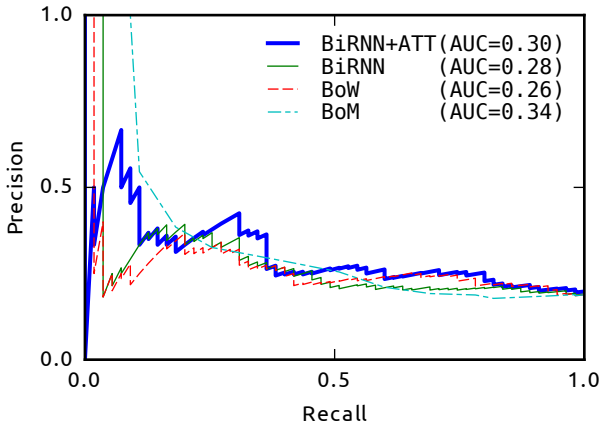
The purpose of this experiment was to test if the proposed RNN with word embedding-based attention model could perform well in a evidence classification task. We benchmarked our method to the RNN without an attention model and conventional lexicon-based classification methods.

3.1 Dataset

We chose seven subject phrases and one or two value phrases for each subject phrase (total of 13 pairs) as shown in Table 1. For each pair of $\mathcal{S}$ and $\mathcal{V}$, we extracted sentences having both $\mathcal{S}$ and $\mathcal{V}$ within two adjacent sentences from Annotated



(a) *Promoting*-claim as positive



(b) *Suppressing*-claim as positive

Figure 2: Precision-recall curves for the classifications of the evidences

English Gigaword (Napoles et al., 2012). From candidates of 7000 sentences, we manually extracted and labeled 1,085 self-contained sentences that support *promoting*/*suppressing* relationship. We allowed sentences in which $\mathcal{S}$, $\mathcal{V}$ did not appear. We chose five subject phrases as training data and other two as test data. Notice that only a fraction of the test data had overlapping value phrases with the training data.

3.2 Metrics

We compared the methods in terms of the area under a precision-recall curve (AUC-PR) because it represents a method's performance well even when data are skewed (Davis and Goadrich, 2006). The area under a curve is obtained by first calculating precision-recall for every possible threshold (precision-recall curve) and integrating the curve with trapezoidal rule. We took the average AUC-PR for when the *promoting* or *suppressing* claim was taken as positive because it was a binary clas-

#		Text
1.	$\mathcal{S}$	Smoking costs some 22 000 Czech citizens their lives every year though the tobacco industry earns huge profits for the nation
	$\mathcal{V}$	Smoking costs some 22 000 Czech citizens their lives every year though the tobacco industry earns huge profits for the nation
2.	$\mathcal{S}$	For the nation the health costs of smoking far outweigh the economic benefits of a thriving tobacco industry the commentary said
	$\mathcal{V}$	For the nation the health costs of smoking far outweigh the economic benefits of a thriving tobacco industry the commentary said

Table 4: Visualization of attention in test data with $\mathcal{S}$ =`smoking` and $\mathcal{V}$ =`income`. Highlights show $\hat{s}_t$ and $\hat{v}_t$. An underlined word had the smallest cosine distance to $\mathcal{S}$ and $\mathcal{V}$, respectively.

sification task. We calculated the area under a receiver operating characteristic curve (AUC-ROC) in a similar manner as a reference.

Since we formulated the learning algorithm in regression-like manner, we chose the cutoff value with the best macro-precision within the training dataset to obtain predicted label. This was used to calculate the macro-precision, the accuracy and the McNemar's test, which were for a reference.

3.3 Baselines

Baselines in this experiment were as follows.

Bag-of-Words (BoW) Dictionary of all words in training/test texts, $\mathcal{S}$ and $\mathcal{V}$ were used. The word counts vector was concatenated with one-hot (or n-hot in case of a phrase) vectors of $\mathcal{S}$ and $\mathcal{V}$ and used as a feature for a classifier.

Bag-of-Means (BoM) The average word embedding (Mikolov et al., 2013) was used as a feature for a classifier.

BiRNN without attention layer This was the same as our method except that it took an average of the BiRNN output and concatenated it with the word vector from $\mathcal{S}$ and $\mathcal{V}$ to be fed into the perceptron; i.e., $\mathbf{z} = \mathbf{x}_s || \mathbf{x}_v || \sum_t (\mathbf{u}_t)$.

We tested BoW and BoM with a linear support vector machine (LSVM) and random forest (RF), and BoW with multinomial naïve bayes (NB). We carried out 5-fold cross validation within a training dataset, treating each subject phrase $\mathcal{S}$ as a fold, to determine the best performing hyperparameters and classifiers. The best performing classifier for BoM was RF with 27 estimators. The best performing classifier for BoW was NB with $\alpha = 0.38$ with no consideration of prior probabilities.

3.4 System setting

We tuned hyperparameters for our method and the BiRNN in the same manner. The best settings are shown in Table 2.

For the BoM, BiRNN and BiRNN+ATT, we used pretrained word embedding of three hundred dimensional vectors trained with the Google News Corpus[1]. We pretrained the BiRNN and the BiRNN+ATT with the Stanford Sentiment Treebank (Socher et al., 2013) by stacking a logistic regression layer on top of a token-wise average pooling of $\mathbf{u}_t$ and by predicting the sentiment polarity of phrases.

For the BiRNN and BiRNN+ATT, the maximum token size was 40, and tokens that overflowed were dropped.

BiRNN and BiRNN+ATT were implemented with TensorFlow (Abadi et al., 2015).

3.5 Results

The average AUC-PRs and reference metrics are shown in Table 3. BiRNN+ATT performed significantly better than baselines with $p = 0.016$ (BiRNN), $p = 1.1 \times 10^{-15}$ (BoM) and $p = 0.010$ (BoW), respectively (McNemar's test). The BiRNN without attention layer was no better than BoW ($p = 0.41$, McNemar's test).

Precision-recall curves of the baselines and our method are shown in Figure 2.

4 Discussion

By extending the neural attention model using a distributed representation of words, we were capable of applying the neural attention model to the evidence classification task with never encountered words. The results implied that it learned *how* to classify evidences that support a relationship of $\mathcal{S}$ and $\mathcal{V}$, rather than the relationship itself.

The attention layer selects which part of the sentence the model uses for classification with magnitudes of $\hat{s}_t$ and $\hat{v}_t$ for each token. We visualize the magnitudes of $\hat{s}_t$ and $\hat{v}_t$ on sentences extracted from a test dataset shown in Table 4.

We observed that the attention layers react to the target phrases' synonyms and their qualifiers. For

[1]The model retrieved from `https://code.google.com/archive/p/word2vec/`

example, the value `income` reacted to the word `profit` in Table 4, #1. The classification result and ground truth were both *promoting*. Generalization to similar words was observed for other words such as `Marijuana` ($\mathcal{S}$ = `cannabis`) and `murder` ($\mathcal{V}$ = `crime`). This implies that the attention layers learned to focus on important phrases, which was the reason why the proposed method outperformed conventional BiRNN without an attention layer.

The method failed in Table 4, #2 in which the ground truth was *suppressing* and the method predicted *promoting*. The method shortsightedly focused on the word `benefits` and failed to comprehend longer context. As a future work, we will incorporate techniques that allow our model to cope with a longer sequence of words.

5 Conclusion

We proposed a RNN with a word embedding-based attention model for classification of evidences. Our method outperformed the RNN without an attention model and other conventional methods in benchmarks. The attention layers learned to focus on important phrases even if words were never encountered, implying that our method learned *how* to classify evidences that support a claim of a relationship of subject and value phrases, rather than the relationship itself.

References

[Abadi et al.2015] Martn Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Man, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Vigas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

[Bahdanau et al.2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the International Conference on Learning Representations (ICLR 2015)*, San Juan, Puerto Rico, May.

[Boltui and najder2014] Filip Boltui and Jan najder. 2014. Back up your Stance: Recognizing Arguments in Online Discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58, Baltimore, Maryland, June. Association for Computational Linguistics.

[Cho et al.2014] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN EncoderDecoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

[Collobert et al.2011] Ronan Collobert, Jason Weston, Lon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural Language Processing (Almost) from Scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November.

[Davis and Goadrich2006] Jesse Davis and Mark Goadrich. 2006. The Relationship Between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA. ACM.

[Duchi et al.2011] John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *J. Mach. Learn. Res.*, 12:2121–2159, July.

[Levy et al.2014] Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context Dependent Claim Detection. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1489–1500, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

[Luong et al.2015] Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal, September. Association for Computational Linguistics.

[Mikolov et al.2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

[Napoles et al.2012] Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated Gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*,

AKBC-WEKEX '12, pages 95–100, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Nguyen and Litman2015] Huy Nguyen and Diane Litman. 2015. Extracting Argument and Domain Words for Identifying Argument Components in Texts. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 22–28, Denver, CO, June. Association for Computational Linguistics.

[Pontiki et al.2015] Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 Task 12: Aspect Based Sentiment Analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado, June. Association for Computational Linguistics.

[Sak et al.2014] Haim Sak, Andrew Senior, and Franoise Beaufays. 2014. Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition. *arXiv:1402.1128 [cs, stat]*, February. arXiv: 1402.1128.

[Sardianos et al.2015] Christos Sardianos, Ioannis Manousos Katakis, Georgios Petasis, and Vangelis Karkaletsis. 2015. Argument Extraction from News. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 56–66, Denver, CO, June. Association for Computational Linguistics.

[Sato et al.2015] Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. 2015. `http://www.aclweb.org/anthology/P15-4019`End-to-end Argument Generation System in Debating. In *Proceedings of ACL-IJCNLP 2015 System Demonstrations*, pages 109–114, Beijing, China, July. Association for Computational Linguistics and The Asian Federation of Natural Language Processing.

[Schuster and Paliwal1997] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45(11):2673–2681.

[Shen et al.2014] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grgoire Mesnil. 2014. Learning Semantic Representations Using Convolutional Neural Networks for Web Search. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14 Companion, pages 373–374, New York, NY, USA. ACM.

[Sobhani et al.2015] Parinaz Sobhani, Diana Inkpen, and Stan Matwin. 2015. From Argumentation Mining to Stance Classification. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 67–77, Denver, CO, June. Association for Computational Linguistics.

[Socher et al.2013] Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y. 2013. Parsing with Compositional Vector Grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465, Sofia, Bulgaria, August. Association for Computational Linguistics.

[Srivastava et al.2014] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

[Yanase et al.2016] Toshihiko Yanase, Kohsuke Yanai, Misa Sato, Toshinori Miyoshi, and Yoshiki Niwa. 2016. bunji at SemEval-2016 Task 5: Neural and Syntactic Models of Entity-Attribute Relationship for Aspect-based Sentiment Analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 294–300, San Diego, California, June. Association for Computational Linguistics.

[Zichao et al.2016] Yang Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, June.