Citation Analysis with Neural Attention Models

Tsendsuren Munkhdalai, John Lalor and **Hong Yu**
University of Massachusetts, MA, USA
tsendsuren.munkhdalai@umassmed.edu, lalor@cs.umass.edu, hong.yu@umassmed.edu

Abstract

Automated citation analysis (ACA) can be important for many applications including author ranking and literature based information retrieval, extraction, summarization and question answering. In this study, we developed a new compositional attention network (CAN) model to integrate local and global attention representations with a hierarchical attention mechanism. Training on a new benchmark corpus we built, our evaluation shows that the CAN model performs consistently well on both citation classification and sentiment analysis tasks.

1 Introduction

Citations are relations between the cited and citing articles and are important content in literature. There are different reasons that authors choose to cite an article. Identifying the purpose of the citations has important applications including faceted navigation, citation based information retrieval, impact factor assessment and summarization of scientific papers (Hearst and Stoica, 2009).

ACA refers to the tasks of citation function classification and citation sentiment analysis. Pioneered by Garfield and others (1965), a large body of citation-related studies have been carried out to develop categorization schemes for citation function analysis. However, most of the studies are limited to to specific domain. The classification schemes are typically complex, containing multiple overlapping categories ranging from three to 35 (Bornmann and Daniel, 2008). In contrast, the success of ACA depends on a small but well-defined set of citation categories. Nanba and Okumura (1999) developed a semi-ACA based on a 3-category scheme derived from Garfield and others (1965)'s 15 categories. Similarly, Pham and Hoffmann (2003) developed rule-based approaches (cue phrases) to classify citations into one of the four classes (*basis*, *support*, *limitation* and *comparison*). Teufel et al. (2009) addressed citation function classification and sentiment analysis jointly by a hierarchical scheme with the top nodes for sentiment and the leaf nodes for function classes. Agarwal et al. (2010) developed a scheme of eight non-overlapping categories for citation function classification in biomedical literatures. This scheme simplifies Yu et al. (2009)'s hierarchical overlapping categories. Recently, a decision-tree based scheme was introduced to facilitate citation context based intelligent systems (Mandya, 2012). The citation function classes, organic and perfunctory proposed by Moravcsik and Murugesan (1975) was adapted for a facet-based classification scheme (Jochim and Schütze, 2012).

Machine learning (ML) approaches to ACA mainly adapted statistical classifiers including support vector machines (SVM), logistic regression and Nave-Bayes classifier (Athar, 2011; Athar and Teufel, 2012; Sula and Miller, 2014). The feature set extracted includes n-grams, part-of-speech tags, word stems, cue phrases, sentence dependency components, named entity mentions and word and sentence location based features. Despite the rich linguistically motivated feature sets, ACA remains a challenge, performing significantly worse than human. One of the reasons for this could be the lack of

Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis (LOUHI), pages 69–77,
Austin, TX, November 5, 2016. ©2016 Association for Computational Linguistics

Category	Description
Function Classification	
Background	Citations that describe background of the main topic on the whole, or provide recent studies and state-of-the-art approaches in a general way
Method	Citations of tools, methods, data and other resources used or adapted in the citing work
Results/findings	Citations that authors used to reference others study to relate their research results and/or findings to the cited work
Don’t know	This category should be chosen if you dont know which one to select
Sentiment Classification	
Negational	Citations that discuss or dispute the correctness and/or weakness of the cited work
Confirmative	Citations that imply to confirm, support or make use of outcomes of the cited work
Neutral	Citations that are not negational nor confirmative
Don’t know	This category should be chosen if you dont know which one to select

Table 1: Citation categories in our analysis scheme.

a large training corpus.

In this study, we report the development of a simplified citation classification schema, a subsequent large annotated corpus, and a deep learning framework for end-to-end ACA.

2 Methods

2.1 Citation Scheme

We developed a simple citation scheme as shown in Table 1. Following Jochim and Schütze (2012), we defined both function classification and sentiment classification schemes as separate facets. For function classification, we followed the widely adopted rhetorical IMARD categories in the scientific domain (Day and Gastel, 2012; Sollaci and Pereira, 2004), and introduced *background*, *method* and *Results/findings* types. We defined the standard *negational*, *confirmative* and *neutral* categories for sentiment classification. We added a *don’t know* category to both function classification and sentiment classification since a previous work shows that such a category improved annotation quality (van Rooyen et al., 2015).

2.2 Machine Learning Approaches

We develop deep neural models and compare them with a baseline model for automated citation analysis.

2.2.1 Long Short Term Memory

Long short-term memories (LSTMs) based models are variations of recurrent neural nets and have been introduced to solve the gradient vanishing problem (Hochreiter, 1998). It has an ability to model long-term dependencies of a word sequence (or context) and has achieved notable success in a varity of NLP tasks like machine translation (Sutskever et al., 2014), speech recognition (Graves et al., 2013) and textual entailment recognition (Bowman et al., 2015). In the context of citation analysis, LSTMs read citation context to construct a dense vector representation of the citation for classification.

Let x_t and h_t be the input and output at time step t. Given sequence of input tokens $x_1, \dots, x_l$ (l is the number of tokens in input text) an LSTM with hidden size k computes a sequence of the output states $h_1, \dots, h_l$ as

$$i_t = \sigma(W_1 x_t + W_2 h_{t-1} + b_1) \quad (1)$$

$$i'_t = \tanh(W_3 x_t + W_4 h_{t-1} + b_2) \quad (2)$$

$$f_t = \sigma(W_5 x_t + W_6 h_{t-1} + b_3) \quad (3)$$

$$o_t = \sigma(W_7 x_t + W_8 h_{t-1} + b_4) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \cdot i'_t \quad (5)$$

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

where $W_1, \dots, W_8 \in R^{k \times k}$ and $b_1, \dots, b_4 \in R^k$ are the training parameters. σ and $\odot$ denote the element-wise sigmoid function and the element-wise vector multiplication. The memory cell c_t and hidden state h_t are updated by reading a word token x_t at a time. The memory cell c_t then learns to remember the contextual information that are relevant to the task. This information is then provided to the hidden state h_t by using a gating mechanism and the last hidden state h_l summarizes the all relevant information. i_t, f_t and o_t are called gates. Their values are defined by non-linear combination of the previous hidden state h_{t-1} and the current input token x_t and range from zero to one. The input gate i_t controls how much information needs to flow into the memory cell while the forget get f_t decides what information needs to be erased in the memory cell. The output o_t finally produces the hidden state for the current input token. The final representation vector h_l is subsequently given to a multi-layer perceptron (MLP) with $softmax$ output layer for classification.

Bi-directional LSTMs read the input sequence in both forward and backward directions and have shown to improve further NLP tasks (Jagannatha and Yu, 2016). We implemented Bi-directional LSTM models for citation classification; here we concatenate the last vector representations of the two LSTMs for the subsequent layers.

Studies have shown that LSTMs based models do not work well on memorizing long sequences (Bahdanau et al., 2015). To overcome this limitation, we introduce the attention models.

2.2.2 Global Attention

Attention mechanisms allow NN models to selectively focus on the most task-relevant part of input sequence. As a result, rather than treating every input vector equally, attention models assign weights to the vectors. Since attention models are able to bring out a past and possibly distant input vector to current time step with the blending operation, it also mitigates the information flow bottleneck in RNNs.

We extend the LSTMs based models with a global attention mechanism. This type of attention mechanism is implemented by a neural network that takes a sequence of vectors (usually output vectors of LSTMs) and selectively blends those vectors into a single attention vector. We adopt the attention architecture proposed by Hermann et al. (2015).

Concretely, the global attention considers all the output vectors $h_1, \ldots, h_l$ to construct an attention weighted representation of the input sequence. Let $S \in R^{k \times l}$ be a matrix of the LSTMs output vectors $h_1, \ldots, h_l$ and $o_l \in R^l$ be a vector of ones. An attention weight vector α, an attention representation r and the final representation h' are defined as

$$M = \tanh(W^a S + W^h h_l \otimes o_l) \tag{7}$$

$$\alpha = softmax(w^\top M) \tag{8}$$

$$r = S\alpha^\top \tag{9}$$

$$h'_l = \tanh(W^s r + W^x h_l) \tag{10}$$

where $W^a, W^h, W^s, W^x \in R^{k \times k}$ are learnable matrices and $w^\top$ is transpose of learnable vector $w \in R^k$. With the outer product $W^h h_l \otimes o_l$ we repeat the transformed vector of h_l l times and then combine the resulting matrix with the projected output vectors.

2.2.3 Compositional Attention Network

The global attention introduced in the previous section does not incorporate subsequence information as it considers the whole input as a single component. However, natural language and its text form are composed of a set of semantic units. For example, a document can be broken down into paragraphs, the paragraphs into sentences, and the sentences into words. Inspired by this, we propose our CAN model. The proposed attention is also hierarchical in a sense that it consists of different attention layers. CAN attends locally over the input subsequences and globally over the whole input and selectively composes these two types of attention representations with a second layer attention to construct a higher level representation. We use the standard neural attention network (Equation (8 - 10)) from the previous section as a main building block in our CAN.

Let $R \in R^{k \times z}$ be a matrix of the representations $r_1, \ldots, r_z$ (z is the number of input subsequences, i.e. the number of sentences in the input) learned by local attentions and r the output of the global attention. We then obtain the final attention representation r' and the final output h'' as follows.

$$M' = \tanh(W'^a S + W'^h r \otimes o_z) \tag{11}$$

$$\alpha' = softmax(w'^\top M') \tag{12}$$

$$r' = R\alpha'^\top \tag{13}$$

$$h'' = \tanh(W'^s r' + W'^x r) \tag{14}$$

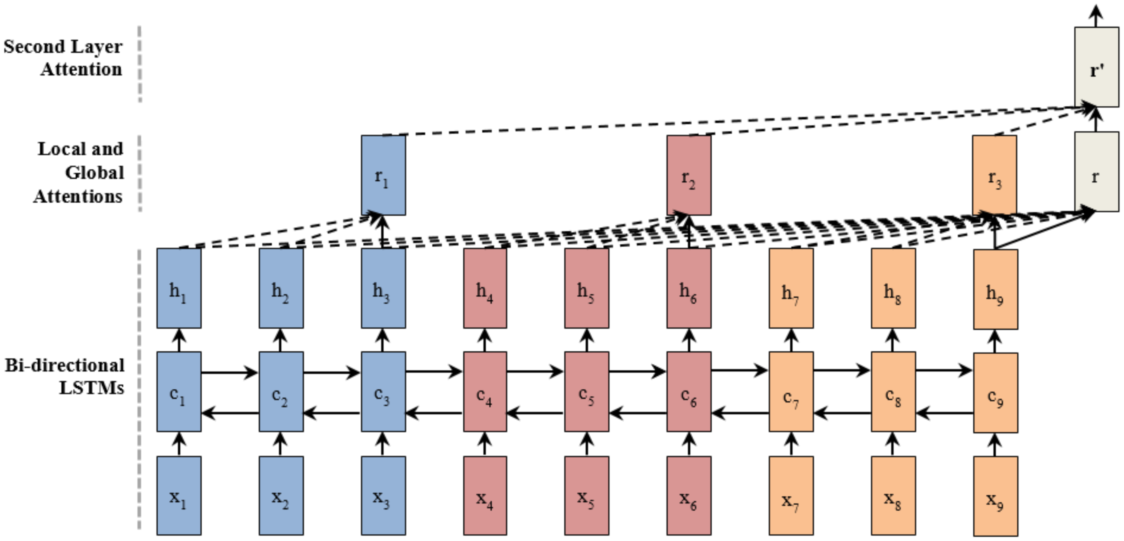

Figure 1: Compositional attention network. r_1, r_2 and r_3 are the locally attended vectors of the output subsets and r is the globally attended vector of the whole output. In the second layer attention, we selectively blend these vectors and obtain the higher level representation r'.

The W matrices and the w vectors of this model can be tied together. When tied, the number of parameters is equal to that of the global attention models. Therefore, this attention network introduces no parametric complexity to compare with the classic global attention model. Figure 1 depicts the overall structure of this model (Equation (1–6), (8–9) and (12–14)). The input consists of the three subsequences $[x_1, x_2, x_3]$, $[x_4, x_5, x_6]$ and $[x_7, x_8, x_9]$. The local attention vectors r_1, r_2 and r_3 are constructed by attending over the LSTM outputs for the each subsequence. Similarly, the global attention vector r is obtained by attending over the whole output sequence $h_1, \ldots, h_9$. In the second layer attention, these representation are composed for the higher level representation r'. The final representation h'' can be obtained according to Equation (9).

The intuition behind our CAN is to attentively compose words within a sentence to construct a local attention vector for each sentence and then these attention vectors are further composed in a second layer attention to learn a whole document representation. We tie the parameters of local, global and the second layer attentions so CAN is forced to learn to compose both the word and sentence presentations attentively.

We also build the bi-directional variation of these models by feeding the concatenated outputs of the forward and backward LSTMs. Due to the concatenated outputs, the size of the W matrices and w vector become $2k \times 2k$ and $2k$ respectively, increasing the number of parameters to be trained.

2.2.4 Baseline Classifier

We implemented a baseline model, which includes extraction of *TF-IDF* statistics of *n-grams* (*1*, *2* and *3-grams*) from each citation for feature sets and a support vector machine (SVM) classifier with a linear kernel. For the SVM model, we performed a grid search over its hyper-parameters (including the regularization parameter, C) by using the development set for evaluation. Once the best parameters were found, the final SVM model was learned on both the training and development sets and tested on the test set.

2.3 Data Collection, AMT Annotation and Gold Standard Datasets

In order to increase the generalization of data, we maximizes the total number of selected articles. Specifically, we selected a total of 5,000 citation sentences from 2,500 randomly selected PubMed Central articles (we randomly selected two citation

Corpus	#docs	Avg. #sents	Max. #sents	#classes	Class Distribution
Yelp 2013	335,018	8.9	151	5	.09/.09/.14/.33/.36
IMDB	348,415	14.02	143	10	.07/.04/.05/.05/.08/.11/.15/.17/.12/.18

Table 2: Statistics for the document-level sentiment datasets.

sentences from each article). We then developed guidelines and deployed an annotation task in a crowdsourcing platform, Amazon Mechanical Turk (AMT).

Each citation was labeled by five annotators. We provide the AMT annotators the previous and the next sentences of the citation sentence to enrich the context. We designed a quality control (attention check questions) and ended the AMT session if the AMT workers failed to answer correctly the attention check questions. To evaluate the quality of annotation, we asked a domain expert (a MD) to independently annotate 100 citation sentences randomly selected from our corpus and used it as the gold standard to evaluate inter-annotator agreement with the AMT workers.

We built two gold standard datasets to use for training and for evaluation. The first dataset is composed of labels agreed by at least three of the five annotators (three label matching). This resulted in 3,422 citations for the function analysis and 3,624 citations for the sentiment analysis. The second dataset is more relaxed in which we selected a label given by the majority of the five annotators. In this setting, we included a label that may fail inclusion by the first approach. For example, even if only two annotators agreed on a label, we will include it in our gold standard dataset because it represents a clear majority vote (the rest of three labels all differ). As a result, this dataset included 4,426 citations for the function classification and 4,423 citations for the sentiment classification.

2.4 CAN for Document-level Sentiment Analysis

In order to test the robustness of the CAN model, we also evaluate it for sentiment analysis on two publically available large-scale datasets: the IMDB movie review and Yelp restaurant review datasets. Particularly, we used the pre-split datasets by Tang et al. (2015). Each document in the datasets is associated with human ratings and we use these ratings as gold labels for sentiment classification. Table 2 reports the statistics for the datasets.

2.5 Experimental Settings

During the experiment, citations labeled with *don't know* were removed from the training data. Each dataset was split into 200/200/rest for dev/test/train sets with a stratified sampling. A stratified sampling is performed to preserve percentage of the citations for each class in each set. We experimented with using only the citation sentence as input example and the expansion with both the previous and the next sentences.

We used ADAM (Kingma and Ba, 2014) for optimization of the neural models. The size of the LSTM hidden units was set to 200. All neural models were regularized by using 20% input and 30% output dropouts and an l_2 regularizer with strength value 1e-3. A word2vec (Mikolov et al., 2013)

Citation Analysis Task	Class	Citation Distribution	
		Majority Voting	Three Label Matching
Function classification	Background	30.5%	20.5%
	Method	23.9%	18.2%
	Results/findings	45.3%	38.3%
	Don't know	0.1%	0.06%
Sentiment classification	Negational	4.8%	2.6%
	Confirmative	75%	59.8%
	Neutral	19.8%	19%
	Don't know	0.2%	0.1%

Table 3: Statistics for our automated citation analysis corpus.

Model	Majority Voting		Three Label Matching	
	Train	Test	Train	Test
SVM	**99.19**	54.27	**87.5**	53.89
LSTMs	59.71	59.55	63.05	66.42
LSTMs + Global Attention	69.02	**65.73**	69.05	**68.61**
Bi-LSTMs	62.14	64.04	67.31	67.88
Bi-LSTMs + Global Attention	72.58	64.6	67.4	**68.61**

Table 4: Citation function classification results. Single **citation sentence** is presented as input.

model trained on a collection of PubMed Central documents transformed citation context to word vectors with size of 200 (Munkhdalai et al., 2015). The parameters of CAN are tied and equal to that of the global attention. The neural models were trained only on the training set while SVM model was built on both training and development sets. We use the development set to evaluate the neural models for each epoch to choose the best model. Each model was given 30 epochs, which was empirically found to be enough time for the models to converge to an optima. The final performances of the methods were reported on the test set. The average training time for the neural network models was approximately three hours on a single GPU (GeForce GTX 980).

3 Results

Table 3 lists the detailed statistics of our AMT annotated corpus. The overall agreement between the expert's annotation and the AMT annotation was 63.1% and 64.7% for function and sentiment analysis tasks. For the function classification, a majority of citations were annotated as results and findings. As shown in Table 3, for the sentiment classification, 4.8% was labeled as *Negational* while 75% and 19.8% were *Confirmative* and *Neutral*. This shows that the citations bias towards a positive statement, resulting a highly unbalanced class distribution.

3.1 Citation Function Analysis

Table 4 lists the results of the function classification by using only citation sentences as input to the models. The SVM baseline obtains the lowest training error. As the models become complex the performance increases. However, some cases like the Bi-LSTMs based global attention model tend to overfit the training data. The unidirectional LSTMs with global attention achieves the best F1-score in both settings when only the citation sentence is input.

Table 5 shows the performance where the inputs are represented by a larger context of the previous, citation and next sentences. We treated the each sentence related to a citation as a subsequence and applied our CAN. Here the bi-directional LSTMs with CAN is the clear winner in terms of the test performance. This model achieves 75.86% F1-score improving the results of the previous model by nearly 7% in the three label matching setup. Unlike the compositional models, the performance of the global attention models decreased in response to additional context given in the input. Furthermore, the models tend to get a higher F1-score in the three label matching setup because this setting has an extra annotation noise filter in selecting the gold labels.

Model	Majority Voting		Three Label Matching	
	Train	Test	Train	Test
SVM	**81.19**	45.72	**99.97**	58.44
LSTMs	53.08	56.74	61.4	59.85
LSTMs + Global Attention	58.97	57.48	77.38	64.96
LSTMs + CAN	60.55	60.11	66.64	73.28
Bi-LSTMs	55.56	56.17	74.49	67.88
Bi-LSTMs + Global Attention	60.28	56.88	66.7	66.42
Bi-LSTMs + CAN	71.34	**60.67**	79.76	**75.57**

Table 5: Citation function classification results. **Citation sentence + its left and right sentences** are used as input.

Model	Majority Voting		Three Label Matching	
	Train	Test	Train	Test
SVM	**82.89**	74.5	**96.39**	70.73
LSTMs	75.25	75.14	73.45	74.48
LSTMs + Global Attention	76.24	76.27	77.38	**75.86**
Bi-LSTMs	75.84	75.7	73.78	75.17
Bi-LSTMs + Global Attention	75.65	**77.4**	80.77	74.48

Table 6: Citation sentiment classification results. Single **citation sentence** is presented as input.

Model	Majority Voting		Three Label Matching	
	Train	Test	Train	Test
SVM	**82.87**	75	**85.72**	71.95
LSTMs	75.25	75.7	73.54	73.79
LSTMs + Global Attention	76.14	75.14	74.68	74.48
LSTMs + CAN	79.18	**76.04**	73.78	**78.1**
Bi-LSTMs	76.24	75.7	73.57	73.79
Bi-LSTMs + Global Attention	75.52	75.7	74.8	74.48
Bi-LSTMs + CAN	75.5	75.44	74.51	75.18

Table 7: Citation sentiment classification results. **Citation sentence + its left and right sentences** are used as input.

3.2 Citation Sentiment Classification

Table 6 shows the evaluation results when the citation sentences are the input. The LSTMs based global attention models obtain the best F1-scores on the test sets. In Table 7, we report the results of the wider context input (citation sentence + its left and right sentences). Here the CAN models perform the best. Similar to the function classification results, the extra context information provides an increasing performance if the model is able to properly exploit.

Despite the same number of training parameters, our compositional attention mechanism significantly improved the performance.

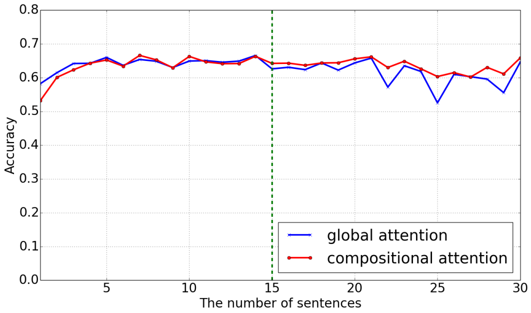

Figure 2: Result on varying length-documents.

3.3 CAN for Document-level Sentiment Analysis

Table 8 lists our document-level sentiment analysis result on the restaurant and movie review datasets. The CAN model achieves a state-of-the-art by locally and globally composing sentences with its hierarchical attention. The Conv-GRNN and LSTM-GRNN are the best-performing models from Tang et al. (2015)'s and are stacked models of convolutional network and RNNs. Our attention models achieve lower MSEs than the stacked models.

We also analyzed whether lengths influence the performance. We split the Yelp dataset into train/dev/test so the models see only documents with length up to 15 sentences during training and classifies much longer documents with length up to 30 sentences during test. Figure 2 plots the test performance over different lengths. The two attention models perform identically on seen lengths except that the global attention model obtains a performance gain on the shorter documents with up to five sentences. However, for unseen lengths (the right side of the green line) the performance of the compositional attention network remains almost consistent and in contrast the global attention starts to decrease in general. This shows the compositional ability of our neural net.

Model	Yelp 2013		IMDB	
	Accuracy	MSE	Accuracy	MSE
SVM (Tang et al., 2015)	59.8	0.68	40.5	3.56
Conv-GRNN (Tang et al., 2015)	63.7	0.56	42.5	**2.71**
LSTM-GRNN (Tang et al., 2015)	**65.1**	**0.5**	**45.3**	3.0
LSTMs + Global Attention (Ours)	63.82	0.57	38.82	**2.25**
LSTMs + CAN (Ours)	**64.49**	**0.55**	**44.16**	2.5

Table 8: Results of document-level sentiment classification. MSE: mean squared error (lower is better).

4 Conclusion

We have developed a generic and simple categorization scheme and a new benchmark corpus for automatic citation analysis. We presented several neural attention networks for the task and evaluated them by using the benchmark corpus. Among these attention mechanisms our original model, we called compositional attention network, performed consistently well on both citation function and citation sentiment classification tasks by attentively composing additional contextual information provided. In an extended experiment, we have also shown that the compositional attention network generalizes better to examples with unseen longer lengths thanks to its compositional operation.

Acknowledgments

We thank the anonymous reviewers for their comments and suggestions. This work was supported in part by the grant HL125089 from the National Institutes of Health (NIH). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

Shashank Agarwal, Lisha Choubey, and Hong Yu. 2010. Automatically classifying the role of citations in biomedical articles. In *Proceedings of American Medical Informatics Association Fall Symposium (AMIA), Washington, DC*, pages 11–15. Citeseer.

Awais Athar and Simone Teufel. 2012. Context-enhanced citation sentiment detection. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 597–601. Association for Computational Linguistics.

Awais Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 student session*, pages 81–87. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Lutz Bornmann and Hans-Dieter Daniel. 2008. What do citation counts measure? a review of studies on citing behavior. *Journal of Documentation*, 64(1):45–80.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642.

Robert Day and Barbara Gastel. 2012. *How to write and publish a scientific paper*. Cambridge University Press.

Eugene Garfield et al. 1965. Can citation indexing be automated. In *Statistical association methods for mechanized documentation, symposium proceedings*, volume 1, pages 189–92. National Bureau of Standards, Miscellaneous Publication 269, Washington, DC.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 6645–6649. IEEE.

Marti A Hearst and Emilia Stoica. 2009. Nlp support for faceted navigation in scholarly collections. In *Proceedings of the 2009 workshop on text and citation analysis for scholarly digital libraries*, pages 62–70. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1684–1692.

Sepp Hochreiter. 1998. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116.

Abhyuday Jagannatha and Hong Yu. 2016. Bidirectional recurrent neural networks for medical event detection in electronic health records. In *NAACL 2016*.

Charles Jochim and Hinrich Schütze. 2012. Towards a generic and flexible citation classifier based on a faceted classification scheme.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Angrosh Annayappan Mandya. 2012. *Enhancing Citation Context based Information Services through Sentence Context Identification*. Ph.D. thesis, University of Otago.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Michael J Moravcsik and Poovanalingam Murugesan. 1975. Some results on the function and quality of citations. *Social studies of science*, 5(1):86–92.

Tsendsuren Munkhdalai, Meijing Li, Khuyagbaatar Batsuren, Hyeon Park, Nak Choi, and Keun Ho Ryu. 2015. Incorporating domain knowledge in chemical and biomedical named entity recognition with word representations. *J. Cheminformatics*, 7(S-1):S9.

Hidetsugu Nanba and Manabu Okumura. 1999. Towards multi-paper summarization using reference information. In *IJCAI*, volume 99, pages 926–931.

Son Bao Pham and Achim Hoffmann. 2003. A new approach for scientific citation classification using cue phrases. In *AI 2003: Advances in Artificial Intelligence*, pages 759–771. Springer.

Luciana B Sollaci and Mauricio G Pereira. 2004. The introduction, methods, results, and discussion (imrad) structure: a fifty-year survey. *Journal of the medical library association*, 92(3):364.

Chris Alen Sula and Matthew Miller. 2014. Citations, contexts, and humanistic discourse: Toward automatic extraction and classification. *Literary and Linguistic Computing*, 29(3):452–464.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Duyu Tang, Bing Qin, and Ting Liu. 2015. Document modeling with gated recurrent neural network for sentiment classification. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1422–1432.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2009. An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*, pages 80–87. Association for Computational Linguistics.

Brendan van Rooyen, Aditya Menon, and Robert C Williamson. 2015. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems*, pages 10–18.

Hong Yu, Shashank Agarwal, and Nadya Frid. 2009. Investigating and annotating the role of citation in biomedical full-text articles. In *Bioinformatics and Biomedicine Workshop, 2009. BIBMW 2009. IEEE International Conference on*, pages 308–313. IEEE.