What can be improved? Identifying actionable items from patient narratives

Lili Shang, Meiyun Zuo*
Research Institute of Smart Senior Care, School of Information, Renmin University of China, Beijing, China
shangll@ruc.edu.cn, zuomy@ruc.edu.cn

2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313550

***Abstract*—Recent literature suggests the potentiality of patient narratives to evaluate the quality of healthcare services, but little research has investigated the "actionability" of such narratives in driving or reshaping the practices. Moreover, the colloquial and implicit nature of natural language expressions poses challenges to automatic *actionable item identification (AII)* from large-scale patient narratives. In this paper, we propose a Content Sense and Inference BiLSTM (CSI-BiLSTM) model to address the challenges, and the model includes the element-aware network (EAN) and the aspect indicator network (AIN). The EAN is designed to sense and capture the mentioned degree of the key elements of input sentences to generate effective element-aware sentence representations. The AIN is endowed with reasoning ability for implicit expressions by inferring whether given opinion words are aspect indicators, to further enhance the accuracy and robustness of the model. To the best of our knowledge, this is the first study to automatically identify actionable items from patients' online narratives. Experiment results on two real-world datasets demonstrate the performance of the proposed model outperforms the competitive baselines and extensive analysis reveals the effectiveness of our model in identifying actionable items.**

***Index Terms*—actionable information, patient reviews, healthcare service improvement, health informatics, natural language processing**

I. Introduction

Nowadays the Internet (e.g., user review websites, physician rating platforms, online health communities) allows patients to write narrative comments about treatment feelings, experiences and views toward their received healthcare services. These narratives contain rich information that not only serves as a reference for other patients to make reasonable choices, but also assists service providers in assessing and enhancing their quality of care. Recognizing the potentiality of this type of patient narratives, researchers from health informatics as well as natural language processing (NLP) domains have put much efforts in analyzing and mining such text data (e.g., [1], [2]).

Despite all this, existing studies have not yet addressed the critical issues regarding how to efficaciously incorporate patients' narratives into practices in a meaningful manner. First, previous research on patient narrative analysis was mainly limited to topic mining [1], [3], and has rarely investigated the "actionability" of such narratives. Second, patients' negative narratives were considered more influential by previous studies [4], [5], while the value of positive narratives was often neglected. Actually, positive narratives serve as valuable means that not only inform particular practitioners to maintain good practices, but also encourage other practitioners to improve practices [6]. Third, the unstructured and implicit nature of natural language expressions makes it challenging to automatically identify actionable items, which means that even current competitive standard deep learning models are no longer so effective. By observing the language expressions of patient narratives, we found that some specific aspects of healthcare services are often not explicitly mentioned but are implied [8]. The vagueness and implicitness feature discussed above makes it more difficult to identify actionable items. It motivates us to realize that, a reasonable and robust model for actionable item identification task should possess the reasoning ability for implicit expressions.

To address the aforementioned challenges, this paper aims to automatically identify actionable items from large-scale patient narratives. In the current study, an *actionable item* is the narrative sentence within which all the key elements of $\langle R, A, P\rangle$ (i.e., roles, aspects and sentiment polarity) are mentioned explicitly or implicitly. We formulate the **actionable item identification (AII)** task as a sentence-level classification problem, which aims at identifying those sentences that are recognized as actionable items. On this basis, we propose a Content Sense and Inference BiLSTM (CSI-BiLSTM) model for the AII task. The model consists of two modules: the element-aware network (EAN) and the aspect indicator network (AIN). The EAN is designed to sense and capture the mentioned degree of the key elements of the input sentences and generate element-aware representations. The AIN is endowed with reasoning ability for implicit expressions by inferring whether given opinion words are aspect indicators, to enhance the accuracy and robustness of the model. Experiments conducted on two datasets demonstrate the performance superiority of the proposed model compared to competitive baselines.

This paper can enrich the literature in health informatics and NLP domains, and the main contributions are summarized as follows:

- To the best of our knowledge, this is the first study to automatically identify actionable items from patients' online narratives. Actionable items identified by our proposed approach can inform relevant healthcare-service practitioners to modify, restrain or reinforce specific actions in their subsequent practices.
- We propose a CSI-BiLSTM model for the AII task, which

*Corresponding Author.

holds the capacity to sense key elements in sentences and identify hidden aspects of healthcare services.

- We evaluate our model on two real-world datasets and experiment results demonstrate that the proposed model outperforms current competitive baselines.

II. Related Work

This research is mainly inspired by two lines of studies: actionable information identification and representation learning. Previous research has not given a careful definition for actionable information, and thus existing methods are not generalized enough [9]–[11]. In the field of health informatics, Grob et al. [6] claimed that researchers ought to identify a pathway to comprehend patient narratives and incorporate these narratives into clinical practices. In their paper, actionability of patient narratives depends to some extent on whether key roles have been mentioned, the information sufficiency of negative narratives, the specificity of positive reviews as well as the universality of the practices illustrated by patients. They further developed a coding template and conducted AII manually. However, this manual method would be impractical when processing large-scale data. Given the above, we argue that a generalized and automatic method for the actionable information identification, especially for patient narratives, is worthy of further exploration.

Representation learning has attracted much attention and achieved great success in NLP [12]. Previous research demonstrates the value of word-level representation and sentence-level representation in enhancing the predictive performance of deep learning models. For word-level representation, semantic and syntactic joint embeddings in a low-dimensional space have been proved to enhance the ability of medical text processing [13] and fine-grained sentiment analysis [14], as it facilitates the capture of features about the linear context and hierarchical context information simultaneously. For sentence-level representation learning, it is usually necessary to associate sentence representations with specific tasks [8], [15], [16].

III. Methods

A. Problem Statement

The objective of the AII task is to extract sentences deemed as actionable items from patient narratives. Based on the analysis of patient narratives and inspired by Grob et al. [6], an *actionable item* is the narrative sentence within which the key elements $\langle R, A, P\rangle$ are all mentioned explicitly or implicitly. Specifically, R represents the role in charge, A means the aspect of the healthcare service and P is the sentiment polarity toward the given aspect. Naturally, a narrative sentence is **actionable**, if it contains all of the key elements $\langle R, A, P\rangle$. In contrast, sentences with key elements missing are **non-actionable**. An example of key elements is shown in Fig. 1. Specifically, sentiment polarity is often reflected by opinion words (i.e., those terms of a sentence used to express attitudes or opinions explicitly) in sentences [7]. Moreover, we call one type of opinion words, which are clues to infer the implicit aspects, as aspect indicators. An aspect indicator would be the equivalent of an aspect in determining the actionability of the narratives.

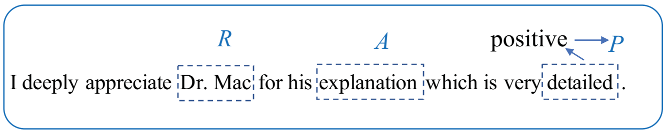

Fig. 1: An Example of the key elements.

B. The Proposed Model

In this study, we formulate the AII task as a sentence-level classification problem. For a given sentence $S = \{w_1, w_2, ..., w_n\}$ with n words, the output is the final predict value y, where $y \in \{0, 1\}$ (actionable: 1, non-actionable: 0). We propose a model called **Content Sense and Inference BiLSTM (CSI-BiLSTM)** for AII, which is summarized in Fig. 2. We will describe module in detail below.

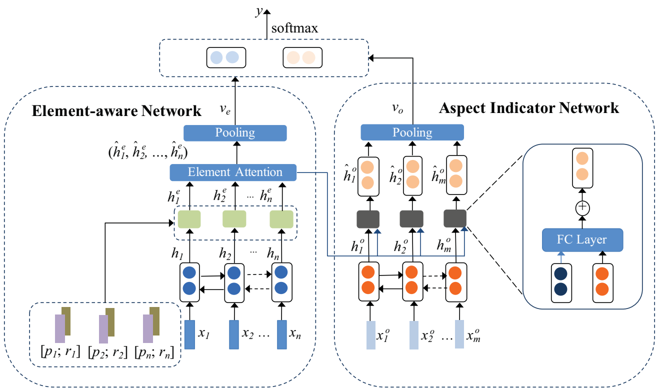

Fig. 2: Overview of the proposed model.

(i) Element-aware network (EAN)

The EAN is designed to sense and capture mentioned degree of the key elements of the input sentences and generate effective element-ware sentence representations. To better represent the key elements, we consider semantic (meaning) features and structure (syntactic information) features. For semantic features, word embeddings are efficient representations for word meanings. For structure features, two types of syntactic knowledge are leveraged: part-of-speech (POS) tags and dependency relations. POS knowledge is important for representing key elements because the part of speech of key elements is often traceable; for example, nouns are often considered as potential aspect words [17]. Moreover, the key elements often present a specific association with each other, such as "opinion words-amod-aspect words", which has been recognized as a common syntactic pattern [18]. Therefore, it is also necessary to include dependency relation knowledge as the input.

First, for each word w_i in the considered sentence, we generate its word representation $x_i \in \mathbb{R}^{dim}$ and the given sentence is denoted as $X = \{x_1, x_2, ..., x_n\}$. In the present study, we adopted bi-directional long short-term memory (BiLSTM) networks [19] to learn sentence representations. Correspondingly, each word's hidden state h_i ($h_i \in \mathbb{R}^{2dim}$) is:

$$h_i = [\overrightarrow{LSTM}(h_{i-1}, x_i); \overleftarrow{LSTM}(h_{i+1}, x_i)] \tag{1}$$

POS tag embedding is a mapping from the POS tag category to a real value vector. The POS tag embedding vector for the given sentence is initialized as *P* = $\{p_1, p_2, ..., p_n\}$, where $p_i \in \mathbb{R}^{dim}$.

We then obtain the triples from the dependency tree of a sentence, which describes the syntactic structure of the sentence. The dependency relation embedding vector is denoted as: *R*=$\{r_1, r_2, ..., r_n\}$, where $r_i \in \mathbb{R}^{dim}$. Based on the modeling of the sequential context and syntactic structure, the initial sentence-representation is obtained. The new hidden state of each word is expressed as:

$$h_i^e = tan(W_a h_i + W_b[p_i; r_i] + b_e) \tag{2}$$

where W_a and W_b are weight matrices, and b_e is bias.

Following this, we propose the element attention and argue that it can guide the model to focus on the element text. In the element attention, we first set the average of the hidden state of the given sentence e_{pool} as the query vector in Eq. (3):

$$e_{pool} = \sum_{i=1}^{n} h_i^e / n \tag{3}$$

Then the score function s_i for each word is calculated based on the word's hidden state h_i^e and e_{pool}. After that, we obtain the attention weight a_i for each word based on the attention score function s_i:

$$s_i^t = tanh(h_i^e W_t e_{pool}^T + b_t) \tag{4}$$

$$\alpha_i = exp(s_i) / \sum_{i=1}^{n} exp(s_i) \tag{5}$$

where W_t, b_t are weight matric and bias, and tanh is a non-linear function.

Finally, the new weighted representation for w_i is generated as follows:

$$\hat{h}_i^e = \sum_{i=1}^{n} \alpha_i h_i^e \tag{6}$$

Hence, the element-ware hidden states for the sentence are $\hat{H}^e$= $\{\hat{h}_1^e, \hat{h}_2^e, ..., \hat{h}_n^e\}$. For the considered sentence, we obtain the final element-ware sentence representation $v_e \in \mathbb{R}^{2dim}$ by averaging the hidden state of each word:

$$v_e = mean\_pool(\hat{h}_1^e, \hat{h}_2^e, ..., \hat{h}_n^e) \tag{7}$$

(ii) Aspect indicator network (AIN)

Aspect, as one of the key elements with considerable power to distinguish actionability of sentences, include explicit aspects and implicit aspects. We design the AIN which processes the reasoning ability for implicit expressions, and the network needs to predict whether an opinion word appearing in the given sentence is an aspect indicator.

We first uniformly extract opinion words from all the input sentences, e.g., extracting "*friendly*" from "*The nurse is very friendly*". Opinion word embeddings $X^o = \{x_1^o, x_2^o, ..., x_m^o\}$, where $x_j^o \in \mathbb{R}^{dim}$, are then generated. Following this, BiLSTM is employed to learn meaningful opinion representations, and the corresponding hidden states for opinion words are $H^o = \{h_1^o, h_2^o, ..., h_m^o\}$, where $h_j^o \in \mathbb{R}^{2dim}$.

Then, we calculate the relation strength between opinion words and potential aspect words. The calculation is based on the hidden state of the opinion word h_j^o and element-ware hidden state $\hat{h}_i^e$ for each word in the given sentence. Accordingly, the hidden state of relation strength is obtained:

$$\tilde{h}_{i,j}^o = RELU(W_E \hat{h}_i^e + W_O h_j^o) \tag{8}$$

where $ReLU$ is the activation function, $W_E \in \mathbb{R}^{2dim}$ and $W_O \in \mathbb{R}^{2dim}$ are parameters for element-aware word representations and opinion representations respectively. Here the $\hat{h}_i^e$ acts as a "filter" for AIN to detect aspect-related features in the given sentence.

After that, the opinion words are checked to see if they are aspect indicators. To distill the aspect indicator information, we calculate the association score $w_{i,j}^o$ between $\hat{h}_i^e$ and each $\tilde{h}_{i,j}^o$ by non-linear functions. The aspect indicator's hidden state is obtained through weighted sum of relation-strength hidden states, which is denoted as $\hat{h}_j^o$.

$$w_{i,j}^o = Softmax(tanh(\hat{h}_i^e W_{b,j} \tilde{h}_{i,j}^o + b_j)) \tag{9}$$

$$\hat{h}_i^o = \sum_{i=1,j=1}^{n,m} w_{i,j}^o \tilde{h}_{i,j}^o \tag{10}$$

Next, the aspect indicator can be represented by v_o, where $v_o \in \mathbb{R}^{2dim}$. Here v_o is the mean value of the hidden state of each word, expressed as follows:

$$v_o = mean\_pool(\hat{h}_1^o, \hat{h}_2^o, ..., \hat{h}_m^o) \tag{11}$$

C. Model Training

Finally, the actionable sentence representation is the concatenation of v_e and v_o and it is fed into the fully-connected layer for AII:

$$P(y) = Softmax(W[v_e; v_o] + b_y) \tag{12}$$

Where W and b_y are parameters of fully-connected layer. The gradient descent method was selected to train the proposed model. The cross-entropy as a loss function is defined as follows:

$$L = -\sum_{x,y \in D} \sum_{c \in C} P(y_c^g) log(P(y_c)) \tag{13}$$

where $P(y)$ is the predicted actionable-item distribution and $P(y_c^g)$ is the gold actionable-item distribution.

IV. EXPERIMENTS

A. Dataset

The data was collected from two websites: (i) the widely used review sites Yelp[1] that offers reviews for a variety of services, one of which is the healthcare service, and (ii) the specialized physician-rating websites RateMDs[2]. Data was pre-processed by removing invalid characters and words were uniformly converted to lowercase. We then used Stanford CoreNLP[3] to parse the POS for each word and the dependencies between words. Considering that there are no available gold standard annotations for opinion words, we therefore extracted opinion words by utilizing MPQA3 [20] so as to check the existence of aspect indicators in given sentences. Finally, a total of 15,876 review sentences were obtained by random sampling for expert annotations to validate the proposed model. The dataset was divided into training data and testing data through performing a stratified sampling split of 0.7 and 0.3. The characteristic statistics of the datasets are presented in Table I.

TABLE I: The characteristic statistics of the datasets.

DateSets	Training		Testing		Total
	Actionable	Non-actionable	Actionable	Non-actionable	
Yelp	1441	4146	667	1728	7982
RateMDs	1822	3704	727	1641	7894

Three researchers in the fields of NLP and health informatics were invited to label the data. To build the classifier and examine the effectiveness of our method on the AII task, sentences were labeled as either “actionable” (1) or “non-actionable” (0).

B. Experiment Settings

To facilitate a better semantic representation for texts, we adopted GloVe [21] to initialize the word embeddings, and set the word embedding size as 300 in line with previous research [22]. For those words that were out of the vocabulary of GloVe, we randomly initialize their embeddings by sampling from the uniform distribution $U(-0.25, 0.25)$. In our experiment, all biases were initialized to zeros, and all the weight matrices were initialized from the uniform distribution $U(-0.2, 0.2)$.

We employed Adam [23] to optimize the parameters of our model for 20 epochs, with the value of momentum set to 0.9. In this study, the dropout strategy [24] was adopted to alleviate the overfitting issue with the dropout rate set to 0.5. This strategy reduces the interaction between hidden nodes, and makes the model more generalized by not relying too much on some local features. Among other settings, the dimension of word hidden states, POS hidden states and dependency relation hidden states were set to 100 and the context window size to 3. Values of these parameters were chosen via 5-fold cross-validation.

[1] http//www.yelp.com
[2] http//www.ratemds.com
[3] https://nlp.stanford.edu/software/lex-parser.shtml

V. RESULTS AND ANALYSIS

A. Result Comparison

Given that little research has examined, we selected those popular and competitive standard deep learning models suitable for NLP tasks as baselines, which are shown below:

Elman-RNN: It is a popularized simple recurrent neural network which is capable of processing language sequences.

LSTM: Long short-term memory (LSTM) networks is recognized as the popular model for NLP tasks, and can solve gradient vanishing issues.

BiLSTM: We implemented BiLSTM networks combined with word embeddings as a robust baseline. BiLSTM allows the model to better capture bi-directional contexts of sentences.

Besides this, ablation experiments were also carried out to investigate the contribution of each module in our model. The following are models applied for the ablation study in our experiments:

Basic-P-D : BiLSTM with pre-trained word embeddings is set as the basic model to learn the sentence representation. POS embeddings are employed in Basic-P-D to learn the tailor-made representation and dependency embeddings are utilized to obtain the syntactic associations between words.

Basic-P-D-A : Basic-P-D-A employs the proposed element attention mechanism which prompts the model to assign high weights to key-elements text when distinguishing actionable items.

OURS : This is the proposed full model and it is used to show the contribution of the aspect indicator module compared to Basic-P-D-A.

Table II presents the performance of the models discussed above, and the best results are in bold. In this experiment, the Precision (Pre), Recall (R) and F1-score (F1) metrics were used to evaluate the performance of the models.

B. Discussion

We derive several findings from the experiment results as detailed below. (i) Our full model achieves the best results, and it outperforms the best baseline model by 5.97% and 3.96% on the F1 measure. (ii) By observing the ablation experiment results, we found that, they further present the advantages of each proposed module in our model. (iii) We observe that the overall experiment results on Yelp are not as satisfactory as that on RateMDs occurring in all of the models. The possible reason behind this is that, the sentence structure

TABLE II: Performance comparison of the models.

Models	Yelp			RateMDs		
	Pre	R	F1	Pre	R	F1
Elman-RNN	74.60	55.47	63.63	83.41	74.69	78.81
LSTM	74.70	56.67	64.45	84.24	74.29	78.95
BiLSTM	77.87	56.92	65.77	86.01	74.58	79.89
Basic- P-D	78.29	58.57	67.01	86.78	74.60	80.23
Basic-P-D-A	82.21	60.19	69.50	88.21	76.92	82.18
OURS	**82.97**	**63.19**	**71.74**	**90.13**	**78.39**	**83.85**

of Yelp reviews is usually more obscure and informal than RateMDs reviews.

C. Element Attention Visualization

We randomly chose two instances from the test data and visualized the attention weights in Fig. 3 to examine whether the proposed element attention mechanism is consistent with intuitive understanding. Note that in Fig. 3, the deeper the color, the higher the assigned weight to the corresponding word. A high weight means more relevance to the prototypes of key elements that helps distinguish sentence categories.

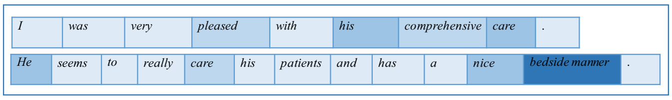

Fig. 3: Visualization of element attention.

We observed that: (i) for the first sentence, higher weights are assigned to "*his*", "*care*" and "*comprehensive*"; (ii) "*He*", "*care*", "*bedside manner*" and "*nice*" are given a relative high level of attention in the second sentence. Obviously, in general, the attention weights assigned by our model to text related to key elements agreed well with human expectation. Above analysis further demonstrates that our model is sensitive to the key elements that help identify actionable items.

VI. Conclusion and Future Work

Actionable items in patient narratives hold great potential to enable healthcare services to be delivered in a way that better meets the needs of patients. However, although AII is of crucial practical and theoretical implications, as yet little has been done to this issue. To this end, this paper gives formal definition for actionable items and proposes a CSI-BiLSTM model for the AII task. To verify the proposed model, a series of experiments were carried out on two datasets from Yelp and RateMDs. Experiment results demonstrate the performance of the proposed model outperforms the competitive baselines and further analysis reveals the effectiveness of our model in identifying actionable items. It is worthwhile for future work to expand the proposed method to enhance the performance of downstream applications such as recommendation systems and reputation management systems.

Acknowledgment

This work was supported by the Key Projects of Philosophy and Social Sciences Research of Chinese Ministry of Education [grant number 19JZD021], the National Natural Science Foundation of China [grant number 71771210], and the Beijing Natural Science Foundation [grant number 9182008].

References

[1] James, T. L., Calderon, E. D., & Cook, D. F. (2017). Exploring patient perceptions of healthcare service quality through analysis of unstructured feedback. Expert Systems With Applications, 479-492.

[2] Lpez, A., Detz, A., Ratanawongsa, N., & Sarkar, U. (2012). What patients say about their doctors online: a qualitative content analysis. Journal of General Internal Medicine, 27(6), 685-692.

[3] Hao, H., & Zhang, K. (2016). The voice of chinese health consumers: A text mining approach to web-based physician reviews. Journal of Medical Internet Research, 18(5).

[4] Liu, J., Hou, S., Evans, R. D., Xia, C., Xia, W., & Ma, J. (2019). What do patients complain about online: A systematic review and taxonomy framework based on patient centeredness. Journal of Medical Internet Research, 21(8).

[5] Zhang, W., Deng, Z., Hong, Z., Evans, R. D., Ma, J., & Zhang, H. (2018). Unhappy patients are not alike: Content analysis of the negative comments from china's good doctor website. Journal of Medical Internet Research, 20(1).

[6] Grob, R., Schlesinger, M., Barre, L. R., Bardach, N. S., Lagu, T., Shaller, D., et al. (2019). What Words Convey: The potential for patient narratives to inform quality improvement. Milbank Quarterly, 97(1), 176-227.

[7] Wu, Z., Zhao, F., Dai, X.-Y., Huang, S., & Chen, J. (2020). Latent Opinions Transfer Network for Target-Oriented Opinion Words Extraction.Proceedings of the AAAI Conference on Artificial Intelligence,34(05), 9298-9305.

[8] Liao, J., Wang, S., & Li, D. (2019). Identification of fact-implied implicit sentiment based on multi-level semantic fused representation. Knowledge Based Systems, 165, 197-207.

[9] Su, P., Mao, W., Zeng, D., & Zhao, H. (2012). Mining actionable behavioral rules. Decision Support Systems.

[10] Kalanat, N., & Khanjari, E. (2019). Extracting actionable knowledge from social networks with node attributes. Expert Systems With Applications.

[11] Mardini, M. T., & Ras, Z. W. (2019). Extraction of actionable knowledge to reduce hospital readmissions through patients personalization. Information Sciences, 1-17.

[12] Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: a review and new perspectives. IEEE Transactions on Software Engineering, 35(8), 1798-1828.

[13] Iyer, M., Zou, C., & Luo, X. (2018). Incorporating syntactic dependencies into semantic word vector model for medical text processing. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE.

[14] Phan, M. H., & Ogunbona, P. O. (2020). Modelling context and syntactical features for aspect-based sentiment analysis. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3211-3220.

[15] Fortuin, V., Weber, R. M., Schriber, S., Wotruba, D., & Gross, M. (2018). InspireMe: learning sequence models for stories. Association for the Advancement of Artificial Intelligence.

[16] Bai, T., Zhang, S., Egleston, B. L., & Vucetic, S. (2018). Interpretable representation learning for healthcare via capturing disease progression through time. Knowledge Discovery and Data Mining.

[17] Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Knowledge Discovery and Data Mining.

[18] Wang, W., & Pan, S. J. (2018). Recursive neural structural correspondence network for cross-domain aspect and opinion co-extraction. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics.

[19] Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing, 45(11), 2673-2681.

[20] Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. International Joint Conference on Natural Language Processing.

[21] Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: global vectors for word representation. Empirical Methods in Natural Language Processing.

[22] Yang, C., Zhang, H., Jiang, B., & Li, K. (2019). Aspect-based sentiment analysis with alternating coattention networks. Information Processing and Management, 56(3), 463-478.

[23] P. Kingma, J. Ba, (2015). Adam: A method for sochastic optimization, in: The 3rd International Conference for Learning Representations.

[24] Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929-1958.