INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Affective Conditioning on Hierarchical Attention Networks applied to Depression Detection from Transcribed Clinical Interviews

Danai Xezonaki[1], *Georgios Paraskevopoulos*[1,3], *Alexandros Potamianos*[1,2,3], *Shrikanth Narayanan*[2,3]

[1]School of ECE, National Technical University of Athens, Athens, Greece
[2]Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, CA, USA
[3]Behavioral Signal Technologies, Los Angeles, CA, USA

dxezonaki@gmail.com, geopar@central.ntua.gr, potam@central.ntua.gr, shri@sipi.usc.edu

Abstract

In this work we propose a machine learning model for depression detection from transcribed clinical interviews. Depression is a mental disorder that impacts not only the subject's mood but also the use of language. To this end we use a Hierarchical Attention Network to classify interviews of depressed subjects. We augment the attention layer of our model with a conditioning mechanism on linguistic features, extracted from affective lexica. Our analysis shows that individuals diagnosed with depression use affective language to a greater extent than not-depressed. Our experiments show that external affective information improves the performance of the proposed architecture in the General Psychotherapy Corpus and the DAIC-WoZ 2017 depression datasets, achieving state-of-the-art 71.6 and 70.3 using the test set, F1-scores respectively.

Index Terms: depression detection, clinical interviews, recurrent neural networks, hierarchical attention networks, affective lexica

1. Introduction

Depression is a serious mood disorder that affects the way people think and behave. According to WHO [1], it is estimated that over 300 million people suffer from depression, which corresponds to the 4.4% of the world's population. Indicative symptoms of depression can be the loss of interest in everyday activities, sleeping and eating disorders, feelings of worthlessness, sadness and exhaustion, or even thoughts of suicide [2]. WHO also states that over 800,000 suicide deaths are reported each year due to depression, while for 15-29-year-old people, it is the leading factor of death. The growing amount of available online data opens opportunities to perform data driven analyses and develop computational algorithms to assist specialists in the field of psychology, study depression and refine clinical methods and protocols.

Depression detection is the problem of identifying signs of depression in individuals. These signs might be identified in peoples' speech, facial expressions and in the use of language. In our task, we consider the binary classification task of detecting depression in transcribed clinical sessions between a therapist and a client. These sessions provide valuable insights of the cognitive and behavioral functioning of clients. Therefore, we leverage behavioral and psycholinguistic cues of the client and therapist language to enhance our models.

Previous studies have shown that depression affects the language use of depressed individuals. They tend to use more absolutist words [3], negatively valenced-words and the pronoun "I" [4] and mention pharmaceutical treatment for depressive disorder [5, 6]. People in distress also make less use of first person plural pronouns [7] and become more self-focused [8]. In [9], linguistic metadata features are employed across with external knowledge including domain-adapted lexica while in [10], Losada et al. propose evaluation methods of existing depression lexica and create sub-lexica based on part-of-speech tagging. Moreover, for the General Psychotherapy Corpus [1], Malandrakis et al. [11] have explored differences in language between therapist and client using psycholinguistic norms and Imel et al. [12] have identified semantic topics discussed in therapy sessions. Other studies based on therapy sessions have also predicted empathy through motivational interviews [13] and have explored behavioral coding learning models for different psychotherapy approaches [14].

Hierarchical models have been proposed for document classification tasks, in order to leverage the hierarchies existing in the document structure and construct a document-level representation based on turn-level and word-level representations [15]. These models have been augmented with attention mechanisms [16, 17] to identify salient words and sentences in the document [18]. In addition, affective lexica have been published [19, 20, 21, 22, 23, 24] which can effectively contribute in sentiment analysis. As a useful external linguistic knowledge, they can be incorporated into neural architectures [25]. In [26], attentional conditioning methods were proven to enhance model performance for sentiment classification tasks.

In this work we focus on the problem of depression detection in psychotherapy sessions. We employ a two-staged hierarchical network functioning at word and turn-level. Each level is equipped with an attention mechanism to extract important content from different parts of the session. To leverage the affective context of depressive language we employ a conditioning method [26] using affective lexica and fuse them in the word-level attention network. We also incorporate the summary attributed to each session into the proposed architectures. Our key contribution is that we integrate existing affective information which improves the results of our hierarchical neural network for depression detection, especially in the case we have small amount of data. This fact results in high performing models and improved robustness across two corpora. We also make our source code publicly available [2].

2. Methodology

Our task is a document classification task, where the input to the model is the transcription of the therapy session and the output

[1]http://alexanderstreet.com
[2]https://github.com/danaiksez/depression-detection

http://dx.doi.org/10.21437/Interspeech.2020-2819

is a prediction of the subjects depression status. Hierarchical Neural Networks are a natural fit for document classification, since sessions are composed of turns, which consist of words, forming a hierarchical textual structure. We further augment the baseline architecture with the integration of external knowledge and the summary provided by the therapist for each session.

2.1. Hierarchical Model

The input sequence of words are embedded into a low-dimensional vector space. In document classification, we want to extract the hierarchies existing in documents in a bottom-up manner. To this end, we use a two-stage hierarchical network that operates at word and turn-level, as we can see in Fig. 1. Both the word-level and the turn-level encoders are implemented using Recurrent Neural Networks (RNN). Since not all words or turns contribute equally to the final session representation, we augment both encoders with an attetion mechanism [16]. At the first level of the hierarchy, a word-level encoder produces turn-level reprepresentations. We feed the words of each turn to the encoder and then combine them to a single representation using an attention mechanism. Let h_{ki} be the annotation of the i-th word in the k-th turn obtained through the word-level encoder. The k-th turn representation results as follows:

$$\begin{aligned} \gamma_{ki} &= g(h_{ki}), \\ \alpha_{ki} &= \frac{e^{\gamma_{ki}}}{\sum_i e^{\gamma_{ki}}}, \\ t_k &= \sum_i \alpha_{ki} \cdot h_{ki} \end{aligned} \tag{1}$$

where g is a learnable mapping, a_{ki} are the attention weights for each word and t_k is the k-th turn representation.

The session representations are extracted in a similar manner. The turn representations t_k are fed into the turn-level encoder and then the attention weights are calculated. The final representations are the weighted sum of the turn-level encoder hidden states with the attention weights.

$$\begin{aligned} \beta_k &= f(t_k), \\ \tau_k &= \frac{e^{\beta_k}}{\sum_i e^{\beta_k}}, \\ r &= \sum_k \tau_k \cdot \beta_k \end{aligned} \tag{2}$$

where f is a learnable mapping, τ_k are the attention weights and r is the session-level representation.

2.2. External knowledge conditioning

According to [4, 10], the affective content can be a distinguishing factor between depressed and not-depressed language. Based on this observation, we employ external linguistic knowledge about the affective content of words. These features can be obtained by sources created by human experts. We consider emotion, sentiment, valence and psycho-linguistic annotations for words. Specifically, we construct a context vector c_{ki} for each word i in turn k, where each dimension corresponds to an annotation from existing affective lexica. We set missing dimensions to zero and we integrate the context vector in the attention mechanism of the word-level encoder. Specifically, we

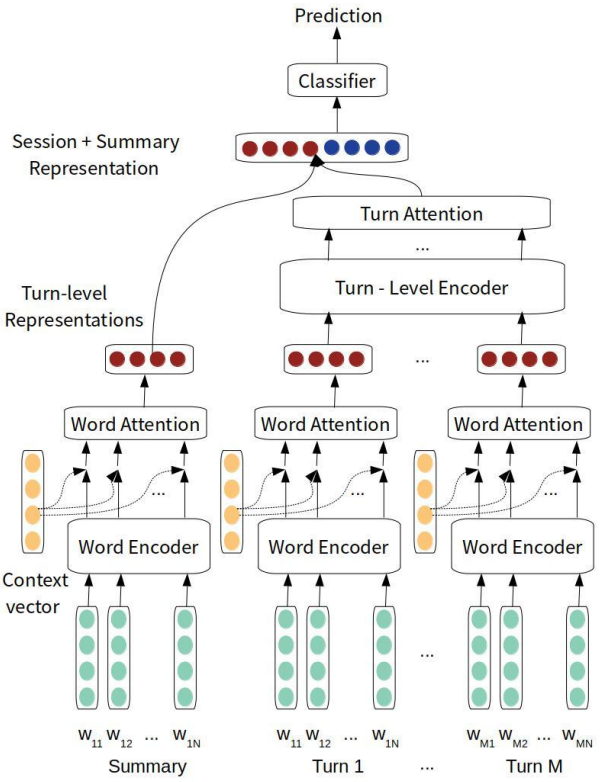

Figure 1: *Hierarchical Model with Attentional Conditioning*

concatentate, $||$, the context vector to the hidden representation of each word h_{ki}, modifying Eq. 1:

$$\begin{aligned} \gamma_{ki} &= g(h_{ki}||c_{ki}), \\ \alpha_{ki} &= \frac{e^{\gamma_{ki}}}{\sum_i e^{\gamma_{ki}}}, \\ t_k &= \sum_i \alpha_{ki} \cdot (h_{ki}||c_{ki}) \end{aligned} \tag{3}$$

Eq. 3 shows that we compute the intermediate representations γ_{ki} using both the word hidden states and the context vector. The softmax function is then applied to γ_{ki} to create the attention weights distribution α_{ki}. The incorporation of external information at this level can force higher values for attention weights corresponding to salient affective words. We also use the concatenated $h_{ki}||c_{ki}$ to create the turn representations t_k to propagate the affective features to the turn-level encoder.

2.3. Integrating Session Summary

In the General Psychotherapy Corpus, each session is accompanied with its summary, denoted as "title", given by an expert. This summary typically consists of 2 or 3 sentences and can be seen as a high-level overview of the topics discussed during the therapy session.

Similar to [27], we extract the summary's vector representation through the word-level encoder and concatenate it directly with the final session representation, before feeding it to the classifier. Let o_t be the summary representation obtained through the word-level encoder. Concatenating it with the session representation (2) produces the final vector $(o_t||r)$ that is

fed to the classifier.

3. Data Overview and Analysis

General Psychotherapy Corpus: We use the *General Psychotherapy Corpus* (GPC) by the "Alexander Street Press". This dataset contains over 1,300 transcribed therapy sessions, which cover a variety of clinical approaches. Metadata are also provided at session level and include demographic information for both therapist and client, the symptoms that the clients are experiencing and a summary attributed to each session, labeled as "title". As some of the sessions are conducted with more than two speakers, we extract a subset of 1,262 sessions which consist of one therapist and one client. Each session is comprised of consecutive dialogue turns, annotated as therapist-side or client-side turns. Among the total of 1,262 sessions, 881 of them are annotated as "not-depressed" samples whereas the rest 381 are annotated as "depressed".

DAIC-WoZ: The DAIC-WoZ dataset is part of the DAIC corpus [28]. It contains a set of clinical interviews which were carried out so as to assist the task of detecting distress disorders. The interview is conducted between a client and a virtual agent serving as the therapist, called Ellie, which is controlled by a human interviewer placed in another location [29]. The dataset contains audio and video recordings as well as the transcripts of clinical interviews. Data are split into train, development and test set, consisting of 107, 35, 47 samples respectively and depression is evaluated on the PHQ-8 depression scale.

3.1. Data Analysis on the General Psychotherapy Corpus

In this section, we explore statistics regarding the language used by depressed and not-depressed individuals in the GPC corpus. As mentioned in Section 2, sessions are provided as consecutive turns, as shown in Fig. 2. We see that the depressed client's language contains more negative affective content. In Table 1, we present the average number of tokens in turns of clients and therapists. We notice that the clients speak twice as much as the therapists on average.

Table 1: *Dialogue turns statistics for therapists and clients*

Features	Sum
Average number of turns/session	196
Average number of tokens in turns	32.3
Average number of tokens in client turns	42.9
Average number of tokens in therapist turns	20.7

Next, we compare the use of language between depressed and not-depressed clients. In particular, we are interested in the use of words that express positive and negative sentiment, sadness and anxiety. To this end, we employ the LIWC lexicon [19], which provides psycho-linguistic annotations for 18,504 words, for 73 different word categories. In Table 2, we compare the vocabularies of depressed and not-depressed people and specifically show the vocabulary sizes and the percentage of their words which are associated with these four affective word categories, in LIWC. We see that depressed subjects use a more consice vocabulary, but include a higher percentage of affective words in it. This hints to the importance of incorporating knowledge about affective language in depression detection.

Moreover, in Table 3, we present the percentages of the four word categories in the language of clients. Specifically, we split the dataset into the samples of depressed and not-depressed people. Subsequently, we use the LIWC lexicon and count the number of occurences of words that belong to each of these affective categories. Finally, we compute their percentages, in the total words of the samples of depressed and not-depressed people. The results indicate that depressed individuals tend to use more negatively-valenced words, which stands in agreement with the related literature [4].

Table 2: *Vocabulary use statistics between the two classes*

Features	Depressed	Not-depressed
Samples	381	881
Total turns	41589	88191
Vocabulary size	16166	23201
Number of affective words	1672	2036
Percentage of affective words	10.34%	8.77%

Table 3: *Percentage of occurences of affective word categories in client language across the two classes*

Categories	Depressed	Not-depressed
Positive sentiment	2.17%	2.26%
Negative sentiment	1.38%	1.30%
Sadness	0.32%	0.32%
Anxiety	0.30%	0.22%

4. Experimental Setup

We employ five network architectures. As a weak baseline model we employ Tf-Idf for feature extraction and an SVM classifier with linear kernel (SVM). Moreover, we develop a Hierarchical Attention-based Network with no external knowledge, which is referred to as HAN. Subsequently, we augment this model with affective conditioning at the attention mechanism, where lexicon annotations are concatenated with word hidden states before the word-level attention layer (HAN+L). We also utilize the session summaries that are provided with the GPC dataset and extend the HAN model with the integration of the summary's representation before the classification layer (HAN+S). Our last model results from the combination of the two previous network architectures (HAN+L+S). As there is no summary assigned to the sessions of the DAIC-WoZ corpus, we evaluate the HAN and HAN+L models on this dataset. For our experiments on GPC we report macro-averaged F1 score due to the class imbalance present in the datasets. This score is calculated using 5-fold cross-validation, where each fold contains an $80\% - 20\%$ train-validation split of the original data. For the DAIC-WoZ corpus we follow the experimental procedure of [30], thus we additionally measure the Unweighted Average Recall (UAR) and present results on the development and test set.

Lexical features: Lexical representations for words are extracted from six affective lexica, namely LIWC [19], Bing Liu Opinion Lexicon [20], AFINN [21], Subjectivity Lexicon (MPQA) [24], SemEval 2015 English Twitter Lexicon (Semeval15) [23] and NRC Emotion Lexicon (Emolex) [22]. AFINN, Semeval15 and Bing Liu provide $1D$ positive/negative sentiment annotations for 6,786, 1,515 and 2,477 words respectively. MPQA provides $4D$ sentiment ratings for 6,886

(a) CLIENT: I don't know. Kind of also I had the feeling like this... this last night when my mother's friend called. Like I was in really bad shape and here I was fooling around, getting myself really in bad shape, taking pills that I shouldn't have been taking, and thereby being completely unresponsive to any of my mother's needs, to anybody else's needs, because I'd managed to get myself so messed up, which in a sense is what my mother's done, but I'd just as soon not interfere.
THERAPIST: Like you suddenly saw yourself walling yourself off from everybody and it looked awfully familiar.
CLIENT: It was almost like I was telling my mother's friend, but of course I didn't, "Don't call me about her. I've got my own problems".

(b) CLIENT: In fact on the other hand, there's a-there can be a quality of being too strong so that you become sort of unfeeling, rigid.
THERAPIST: So self contained almost that nobody ever gets in that there aren't any.
CLIENT: That you don't need anybody and you don't-as a result nobody needs or wants you.
THERAPIST: I guess sometimes that looks pretty good momentarily but really when you look at it, you don't want that. Like you don't want to be walking in without people.
CLIENT: Right. I don't know, I think it's probably something I'm just going to have to experiment with, recognizing those alternatives.

Figure 2: *Example of sessions for (a) a depressed client and (b) a not-depressed client. Blue: positive words, Red: negative words, as found in LIWC, AFINN and Bing Liu Opinion Lexicon*

words. Emolex provides $19D$ emotion ratings for 14,182 words. LIWC provides $73D$ psycholinguistic annotations for 18,504 words. The combined six lexica provide a vocabulary coverage of 25,534 words. These features are concatenated into a 99-dimensional context vector.

Data preprocessing: To preprocess the data of the GPC, we keep only the dialogue turns of the therapist and the client by removing speaker tags and any extra information provided as notes in the transcript. Next, for both corpora we tokenize the speaker turns by splitting them into words. We use $300D$ GloVe [31] pretrained word embeddings, trained on the Common Crawl corpus, to extract word representations.

Implementation details: Our model consist ofs two encoder layers, where a Bi-directional Gated Recurrent Unit (GRU) is implemented on the first stage and two more on the second. All encoders use 300 hidden size and 0.2 dropout rate. Model parameters are optimized using Adam with 10^{-3} learning rate. The model is trained for a maximum of 40 epochs and we use early stopping to select the model with the lowest validation loss. For the system implementation, Pytorch framework [32] is used.

5. Results and Discussion

We compare the performance of the proposed models when given as input the client turns (Client), the therapist turns (Therapist) or the whole dialogue (Client+Therapist). In Table 4 we present the results for the GPC dataset. We see that the integration of external affective and psycholinguistic features improves model performance for all model configurations over the HAN and SVM baselines. Furthermore, we notice that when we add the summary information we also gain a performance boost, sometimes greater that the external affective information. Summary and lexica integration leads to a performance increase when we provide only the client data. In addition, we see that the client turns are more important for depression detection, as expected, and incorporation of the therapist turns contributes little to the overall model performance. Based on our results, the best performance can be achieved by the HAN+L+S model, while HAN+L model performs best if such annotation is not available.

In Table 5 we present results for the DAIC-WoZ dataset. We observe that affective conditioning significantly improves the performance over the baseline model (HAN). Our HAN+L model also shows improved F1 and UAR scores over the models proposed in [30]. Overall, we see that conditioning of external psycholinguistic knowledge in this small dataset (189 samples) enhances the performance and the results are comparable to these of the GPC corpus.

Table 4: *Results of different architectures on the GPC*

Experiment	Client	Therapist	Client+Therapist
SVM	0.478	0.464	0.484
HAN	0.681	0.647	0.695
HAN+S	0.698	0.641	**0.718**
HAN+L	0.693	**0.659**	0.706
HAN+L+S	**0.715**	0.640	**0.716**

Table 5: *Results of the DAIC-WoZ corpus*

Method	Devel. Set		Test Set	
	F1-macro	UAR	F1-macro	UAR
[30] HCAN	0.51	0.54	0.63	0.66
[30] HLGAN	0.60	0.60	0.35	0.33
HAN	0.46	0.48	0.62	0.63
HAN+L	**0.62**	**0.63**	**0.70**	**0.70**

6. Conclusions and Future Work

We propose a novel model for depression detection with integrated external affective and psycholinguistic information. Our model is a Hierarchical Attention Network that encodes words and dialogue turns in different levels of the architecture. The external features are integrated into the attention mechanism, forcing the attention weights to focus on salient affective information. The external knowledge integration leads to high performing models and increased robustness for both the small datasets (1262 and 189 samples respectively) we explore. In the future, we plan to extend our architecture to model the dialogue interaction of the therapist and the client. Finally, we plan to incorporate more elaborate information sources, e.g. expert knowledge bases from psychologists.

7. Acknowledgements

We would like to thank psychologists Evangelia Prassopoulou and Anastasios Panopoulos who gave us a thorough insight into Depressive disorder and their psychological approach on the treatment.

8. References

[1] W. H. Organization, *Depression and Other Common Mental Disorders: Global Health Estimates.*, 2017.

[2] *Diagnostic and statistical manual of mental disorders (5th ed.).* American Psychiatric Association, 2013.

[3] M. Al-Mosaiwi and T. Johnstone, *In an Absolute State: Elevated*

Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation, 2018. [Online]. Available: https://doi.org/10.1177/2167702617747074

[4] S. Rude, E.-M. Gortner, and J. Pennebaker, *Language use of depressed and depression-vulnerable college students*. Journal Cognition and Emotion, Pages 1121-1133, 2004. [Online]. Available: https://doi.org/10.1080/02699930441000030

[5] M. Gamon, M. Choudhury, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Association for the Advancement of Artificial Intelligence*, 2013.

[6] N. Ramírez-Esparza, C. Chung, E. Kacewicz, and J. Pennebaker, "The psychology of word use in depression forums in english and in spanish: Testing two text analytic approaches," 01 2008.

[7] "The way we refer to ourselves reflects how we relate to others: Associations between first-person pronoun use and interpersonal problems," *Journal of Research in Personality*, vol. 47, no. 3, pp. 218 – 225, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0092656613000160

[8] N. Mor and J. Winquist, "Self-focused attention and negative affect: A meta-analysis," *Psychological bulletin*, vol. 128, pp. 638–62, 2002.

[9] A. Perušić, D. Kustura, and I. Matak, "Using linguistic metadata for early depression detection in social media," 2018.

[10] D. Losada and P. Gamallo, "Evaluating and improving lexical resources for detecting signs of depression in text," *Language Resources and Evaluation*, pp. 1–24, 08 2018.

[11] N. Malandrakis and S. S. Narayanan, "Therapy language analysis using automatically generated psycholinguistic norms," in *INTERSPEECH*, 2015, pp. 1952–1956.

[12] Z. E. Imel, M. Steyvers, and D. C. Atkins, "Computational psychotherapy research: scaling up the evaluation of patient-provider interactions." *Psychotherapy*, vol. 52 1, pp. 19–30, 2015.

[13] J. Gibson, N. Malandrakis, F. Romero, D. C. Atkins, and S. S. Narayanan, "Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms," in *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association*, pp. 1947–1951.

[14] J. Gibson, D. Atkins, T. Creed, Z. Imel, P. Georgiou, and S. Narayanan, "Multi-label multi-task deep learning for behavioral coding," *IEEE Transactions on Affective Computing*, 2019.

[15] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015, pp. 1422–1432.

[16] D. Bahdanau, K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate*, 2015.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, ukasz Kaiser, and I. Polosukhin, *Attention is all you need*, 2017.

[18] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2016, pp. 1480–1489.

[19] Y. Tausczik and J. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, vol. 29, pp. 24–54, 2010.

[20] M. Hu and B. Liu, "Mining and summarizing customer reviews," 08 2004, pp. 168–177.

[21] F. Årup Nielsen, "A new anew: evaluation of a word list for sentiment analysis in microblogs," 2011, pp. 93–98.

[22] S. Mohammad and P. Turney, "Crowdsourcing a word-emotion association lexicon," *Computational Intelligence*, vol. 29, 2013.

[23] S. Kiritchenko, X. Zhu, and S. Mohammad, "Sentiment analysis of short informal text," *The Journal of Artificial Intelligence Research (JAIR)*, vol. 50, 2014.

[24] T. Wilson, P. Hoffmann, S. Somasundaran, J. Kessler, J. Wiebe, Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, "OpinionFinder: A system for subjectivity analysis," in *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*. Association for Computational Linguistics, 2005, pp. 34–35.

[25] M. Trotzek, S. Koitka, and C. Friedrich, "Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, pp. 588–601, 2018.

[26] K. Margatina, C. Baziotis, and A. Potamianos, "Attention-based conditioning methods for external knowledge integration," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019, pp. 3944–3951.

[27] F. Aris, C. Baziotis, D. Pappas, H. Papageorgiou, and A. Potamianos, "Hierarchical bidirectional attention-based rnn in biocreative vi precision medicine track, document triage task," 2017.

[28] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, S. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), 2014, pp. 3123–3128. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2014/pdf/508_Paper.pdf

[29] D. DeVault, R. Artstein, G. Benn, T. Dey, E. Fast, A. Gainer, K. Georgila, J. Gratch, A. Hartholt, M. Lhommet, G. Lucas, S. Marsella, F. Morbini, A. Nazarian, S. Scherer, G. Stratou, A. Suri, D. Traum, R. Wood, Y. Xu, A. Rizzo, and L.-P. Morency, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, ser. AAMAS '14. International Foundation for Autonomous Agents and Multiagent Systems, 2014, p. 1061–1068.

[30] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. W. Schuller, "A Hierarchical Attention Network-Based Approach for Depression Detection from Transcribed Clinical Interviews," in *Proc. Interspeech 2019*, 2019, pp. 221–225.

[31] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2014, pp. 1532–1543.

[32] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, *Automatic differentiation in pytorch*, 2017.