CONTEXT MODELLING USING HIERARCHICAL ATTENTION NETWORKS FOR SENTIMENT AND SELF-ASSESSED EMOTION DETECTION IN SPOKEN NARRATIVES

Lukas Stappen[1]*, Nicholas Cummins*[1]*, Eva-Maria Meßner*[2]*, Harald Baumeister*[2]*,*
Judith Dineley[1]*, Björn Schuller*[1,3]

[1]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[2]Clinical Psychology and Psychotherapy, University of Ulm, Germany
[3]GLAM – Group on Language, Audio & Music, Imperial College London, UK

stappen@ieee.org

ABSTRACT

Automatic detection of sentiment and affect in personal narratives through word usage has the potential to assist in the automated detection of change in psychotherapy. Such a tool could, for instance, provide an efficient, objective measure of the time a person has been in a positive or negative state-of-mind. Towards this goal, we propose and develop a hierarchical attention model for the tasks of sentiment (positive and negative) and self-assessed affect detection in transcripts of personal narratives. We also perform a qualitative analysis of the word attentions learnt by our sentiment analysis model. In a key result, our attention model achieved an unweighted average recall (UAR) of 91.0 % in a binary sentiment detection task on the test partition of the Ulm State-of-Mind in Speech (USoMS) corpus. We also achieved UARs of 73.7 % and 68.6 % in the 3-class tasks of arousal and valence detection respectively. Finally, our qualitative analysis associates colloquial reinforcements with positive sentiments, and uncertain phrasing with negative sentiments.

Index Terms— state-of-mind, mood congruency, attention mechanisms, hierarchical models, gated recurrent units

1. INTRODUCTION

An individual's current emotional state, as expressed by self-reported valence and arousal, affects their perception, cognition, attention and memory retrieval [1]. The interaction between current emotional state and mental functioning is herein referred to as *state-of-mind* (SoM). The interaction of state-of-mind and mental functioning is mood congruent, thus a positive state-of-mind is shifts attention towards positive cues and vice versa [2]. The same holds for negative emotions and cues. In psychotherapy, this effect can be used to enhance emotion regulation skills, for example by encouraging patients to construct positive narratives about themselves [3, 4].

The potential of personal storytelling in therapeutic settings is strongly supported by recently published results indicating that the telling of personal narratives directly influences SoM [4]. Results presented in the same work also indicate that the sentiment (positive or negative) of such narratives can be determined from word use alone. These results, however, were not obtained using state-of-the-art machine learning methodologies which could enable more efficient and objective analyses. Therefore, to fully realise the potential of these findings, there is a need to assess the efficacy of contemporary learning methods for analysing text for such tasks.

In this regard, we herein propose and develop hierarchical attention networks for the two main tasks of (i) sentiment analysis of personal narratives, and (ii) the prediction of self-assessed emotion related to personal narratives. Neural network based approaches, particularly *Recurrent Neural Networks* (RNN), have been shown to be suitable in related tasks [5, 6, 7, 8]. However, each narrative requiring analysis is approximately 5 minutes long and has only one label [4, 9], and conventional recurrent approaches struggle in such learning conditions [10, 11]. One solution to this problem is to incorporate attention mechanisms [12, 13], specifically hierarchical attention mechanisms [10], into our developed model. Our network can then explicitly model the contribution of each word in a particular sentence towards the target class, as well as modelling the task-specific context at semantically higher levels, such as at the sentence or document level [10]. Attention mechanisms have been used in related tasks such as document-level sentiment analysis [14], and emotion detection from closed captions [15].

The main tasks of this study are sentiment analysis and affect detection in which a plethora of linguistics based approaches have been proposed and developed, e. g. , [16, 17, 18, 19]. It uses the *Ulm State-of-Mind in Speech* (USoMS) corpus [4, 9] and is related to works in the Self-Assessed Affect Sub-Challenge of the Interspeech 2018 Computational Paralinguistics Challenge (COMPARE) [9]. Participants in the challenge were instructed to predict self-assessed valence using *acoustic* feature representations only. *Unweighted Average Recall* (UAR) scores on the challenge test set ranged from 48.9 % through to 68.4 % for a 3-class detection task [20, 21]. The work present herein differs from these approaches by automatically analysing the linguistic content of the USoMS files. To the authors' knowledge, this is the first time such a study has been conducted on the corpus.

2. EXPERIMENTAL CORPUS

All experimental results were obtained using the USoMS corpus [4, 9]. The corpus contains audio recordings and transcriptions of 100 German-speaking participants, recalling two negative and two positive experiences in an interview setting. The purpose of the original study was to capture the transition from one emotional state to another [4]. The participants self-assessed their emotional state before the interview (ground truth) and after each question on a 10-point Likert scale of valence and arousal scores. For classification purposes, these scores are mapped into 3 classes: low, 0-4; medium, 5-7; and high: 8-10.

The study protocol was as follows: **k0** – Self-assessed affect-0, herein referred to as *Ground Truth* (GT); **k1** – *Negative Narrative*-1 (NN-1) and Self-assessed affect-1; **k2** – *Negative Narrative*-2 (NN-2) and Self-assessed affect-2; **k3** – *Positive Narrative*-1 (PN-1) and Self-assessed affect-3; and **k4** – *Positive Narrative*-2 (PN-2) and Self-assessed affect-4. Each question was answered freely in spoken language with a personal narrative approximately 5-minutes long. The recorded sessions were then manually transcribed.

We use our proposed Hierarchical Attention approach (cf. Section 3) to answer three specific research questions on the USoMS data. First, is it possible to classify the binary sentiment (positive or negative) of the narratives? Second, can the *transition* of the SoM through the study protocol be modelled? This particular task is non-trivial, as the affect labels are self-assessed and we cannot assume a common score scale between the participants. This differs from most conventional affect detection tasks in which ground-truth scores are derived from multiple external annotators, e. g. , [22]. Thirdly and finally, can context-dependent, semantically meaningful (concerning the sentiment of the narrative) words in transcriptions of psychotherapy sessions be automatically identified?

3. HIERARCHICAL ATTENTION NETWORK FOR SENTIMENT AND SELF-ASSESSED EMOTION

Our proposed network is based on the hierarchical attention architecture for document classification proposed in [10]; it takes all words of one sentence as the input and passes it through the hierarchical structure (cf. Figure 1). An embedding layer transforms the words w into a d-dimensional word vector x. Then, in hierarchical manner, the word encoder g_w (with parameters H_w) and an attention mechanism α_w (with parameters A_w) transforms the input sequence x, with the maximum length of L words, into a higher-level sentence representation S. This sentence representation is then further compressed into a document representation by a sentence encoder g_s (with parameters H_s) and an additional attention mechanism α_s (with parameters A_s). In our case, a document indicates an answer to the interviewer's question q. This process is repeated for all sentences with the maximum number of T sentences for a question, such that the entire data set can be represented as $D = \{(x_i, y_i), i = 1, ..., N)\}$, where N is the total number of questions and x_i the embedded words $\{w_{11}, w_{12}, ...w_{LT}\}$ with labels y_i.

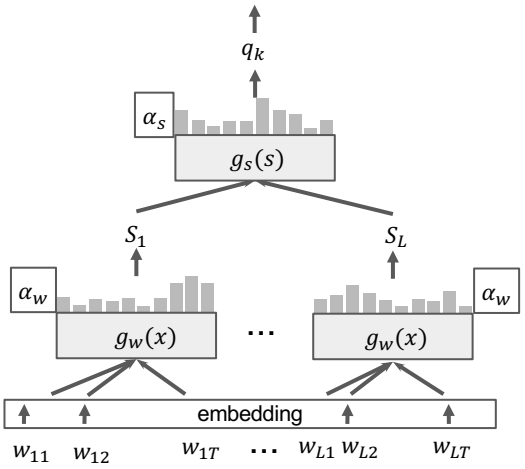

Fig. 1. Basic hierarchical attention neural network architecture. Data is compressed stepwise from bottom to top, starting with words of each sentence, where the word encoder and attention layer learn the most important words in context and merge into sentence representations. The same procedure is then followed for the transformation from sentence to question representation.

3.1. Encoder Layers

The purpose of the encoder layers g with the encoder parameters H_w and H_s is to summarise sequential information i into a meaningful hidden representation h. Alongside the usual encoder compression through a *fully connected feed-forward* (FC) layer, a *Gated Recurrent Unit* (GRU) layer is included to learn the temporal dependencies in the input sequences [23]. Both learn the hidden representation for each level:

$$h_w^{(it)} = g_w(x_l^{(it)}), l \in \{1, ..., L\} \quad (1)$$

for the word encoder, and:

$$h_s^{(i)} = g_s(s_t^{(i)}), t \in \{1, ..., T\} \quad (2)$$

for the sentence encoder.

3.2. Attention Layers

To obtain a more meaningful high-level context vector, we use soft attention mechanisms to emphasise important hidden state vectors of an input sequence. In this regard, word attention emphasises words in a specific context that are significant to the sentence meaning. Sentence attention stresses sentence representations that contribute more to the question representation and thus the prediction quality. To ensure we have attention for every word and sentence, we apply it to all hidden states, not just the final state:

$$u^{(it)} = tanh(W_w h_w^{(it)} + b_w) \quad (3)$$

$$\alpha_w^{(it)} = softmax(u^{(it)T} u_w) \quad (4)$$

$$s_i = \sum_{t=1}^{T} \alpha_w^{(it)} h_w^{(it)}, \quad (5)$$

where $u^{(it)}$ denotes the hidden presentation of $h_w^{(it)}$ by feeding it into a FC with a $tanh$ activation function, the word-level context vector u_w and the importance weight normalized by a softmax function $\alpha_w^{(it)}$. The mechanism works in the same way for sentence attention, where $h_w^{(it)}$ is analogous to $h_s^{(i)}$, u_s is the sentence-level context vector and q represents the question vector summarising all information.

3.3. Question Sequence-to-Sequence Layer

To examine the transition of the SoM in the USoMS corpus (cf. Section 2), the architecture is simply extended by another one-directional sequence-to-sequence layer (QSL) g_t (with the parameter H_t) on top of the existing architecture to predict the participant-dependent, self-assessed labels in sequence:

$$q'_k = \overrightarrow{GRU}(q_k), k \in \{1, ..., 4\}, \quad (6)$$

where k is the question number from one participant P and q'_k is the final question representation, combining the original question representation q_k (in recorded order) and the hidden representation of previous questions $h_{k-1}^{(i)}$. In this case, all components of the base architecture are extended by k, so that, for example, the input becomes $x_i = w_{111}, w_{112}, ..., w_{KTL}$. In our initial exploratory analysis of the data set, we observed that the self-assessed labels are very inconsistent between the participants due to self-assessment. Therefore, we integrated the GT (y_0) for system calibration by concatenating the hot-encoded y_0 with the first question representation q_1 (cf. Figure 2) to learn, the shifts in the SoM between the questions, and not just the label itself. To obtain the fixed input sequence required for GRU processing, a zero vector of the same length is concatenated to $q_k \in \{2, 3, 4\}$.

Until the QSL receives the input q, all four hierarchical models (one per question representation K) run in parallel, whereby parameters of all hierarchical components (H_w, H_s, A_w, A_s) on the same level can be shared to learn a joint representation and accelerate the training process.

3.4. Classification Layers

The high-level representation of the questions q or q' are then used as the input for the final classification layers:

$$\hat{y_k} = \varphi(W_{kc} q'_k + b_{kc}), \quad (7)$$

where W_{kc} is the weight matrix and b_{kc} is the bias. q'_k can also be q_k if no g_t is added to the network and φ is either a sigmoid function if it is a binary with $c = 2$ or a softmax function if is a multi-class $c > 2$ classification. As an objective function, we use the differentiable cross-entropy function between the correct y_i and predicted $\hat{y_i}$ label so that any Stochastic Gradient Decent optimizer can minimise the loss.

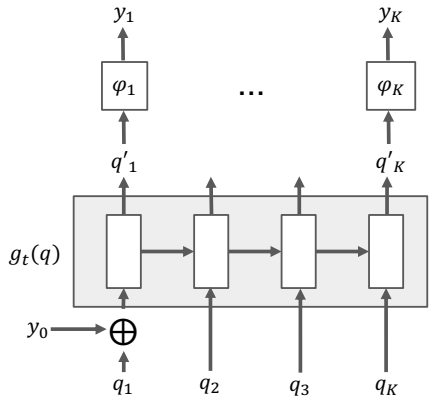

Fig. 2. An illustrative overview of the question sequence-to-sequence layer, which is calibrated with the y_0 ground truth label

4. EXPERIMENTS AND RESULTS

4.1. Data Partitioning

We are using the predefined training, validation (devel) and test partitions of the challenge dataset and refer to [9] for full details. Training and hyperparameter tuning utilise the training and devel set, whereby the test set is exclusively for evaluation.

4.2. Key Settings

We first separate each file into individual sentences. Besides the natural end of sentences, we also used conjunction words such as *"und"* (engl. and) to separate sentences [24]. This parsing had the effect of reducing the input sequences into manageable lengths, resulting in overall 400 questions and 14,333 sentences. As GRUs require a fixed input length, we fixed L to 65 words per sentence and T to 35 sentences per question. We zero padded sequences that do not reach these maximums. All words were vectorised into a 40-dimensional representation using pre-trained German word embeddings [25].

Our models are implemented using Keras customisable layers and Tensorflow. For training, we used the Adam optimiser with a learning rate of 0.001, a clip value of 0.5 and set our mini-batch size to 20. Besides GRUs, we also used *bidirectional GRU* (BiGRU), we also explore the performance of three different encoder types (FC, GRU, BiGRU) in combination with *Attention* (Att). The encoders can either *share encoder weights* (SE), *attention weights* (SA) or both (SEA).

Table 1. Results of 2-class positive and negative sentiment detection on the devel(opment) and test set of the USoMS corpus. Results report in unweighted average recall [%].

Encoder	Non-Attention devel / test	Attention devel / test
FC	77.6 / 81.4	83.6 / 85.2
GRU	82.8 / 86.5	84.5 / 88.5
BiGRU	83.6 / 87.8	86.2 / 91.0

Table 2. Results on the devel(opment) and test set of the USoMS corpus when predicting the transition of self-assessed labels with our sequence-to-sequence question layer. Results report in unweighted average recall [%].

Configuration	Arousal devel / test	Valence devel / test
AttFC-SEA	67.2 / 73.7	63.8 / 65.5
AttGRU-SEA	63.8 / 70.5	64.7 / 68.6
AttBiGRU-SE	68.1 / 70.5	63.8 / 67.3

4.3. Results

4.3.1. Quantitative Results

For the task of predicting the **sentiment** of the narratives, we observed that regardless of the encoder, we achieved better results when using attention. In this regard, we achieved the most apparent result increasing with the simpler FC encoder. In contrast to GRU encoder, which automatically learns the information from previous sequences through their recurrent construction, the FC benefits from the attention context information. However, the results of GRU based encoders are considerably superior, with the strongest test UAR, 91.0 % achieved by the attention enhanced AttBiGRU (cf. Table 1).

For the more complex task of prediction the **self-assessed labels**, using our QSL architecture and infusing y_0 for calibrations, we achieved initial UARs of 64.7 % for Valence and 68.6 % for Arousal. We increased these results to 68.6 % (V) and 73.7 % (A) by combining shared encoder and attention weights (cf. Table 2). Furthermore, we observed similar results by using 0.4 dropout in the initial QSL setup. We speculate this is due to both dropout and the partial sharing of weights having a slight regularisation effect.

Besides, we evaluate our models without QSL to provide approximate comparability with the COMPARE-2018 papers in which the ground truth could not be used. An exact one-to-one comparison is not possible as within the challenge, the audio data was segmented into 8-second chunks [9]. Our results indicate that, surprisingly, AttFC achieved the best results with AttFC-SA of 69.2 % for Arousal and with AttFC-SE of 68.0 % for Valence. The result for valence is only slightly below the best challenge result.

4.3.2. Qualitative Analysis

A qualitative analysis of the context-based word and sentence attentions from predicted participant responses (on the test set), reveals a wide variety in the expression of emotions in the narratives. Besides the identification of already known emotional signal words related to fear, pride or joy, the network learnt less obvious word combinations, which would have been undetectable using conventional methods.

We observed that colloquial reinforcement was highly relevant. For instance, the German *"richtig"* (engl. correct but colloquially can also translate to really or very) in combination with words such as *"spontan"* (engl. spontaneous), *"fertig"* (engl. finished, however colloquial uses include *"richtig fertig"* – engl. really exhausted; or *"richtig wichtig"* – engl. very important), have a strong positive influence. We observed high levels of phrasing associated with uncertainty in the negative case, as initiated by words such as *"irgendwie"*, *"irgendwann"* or *"irgendein"* (engl. somehow, sometime and any). These were often used in conjunction with a verb or noun, for instance, somehow listless or somehow the feeling. The variety of recognised indicators indicate that the method presented is potentially qualitative superior to the context-free word counts of word categories currently used for psychological data sets and should be further explored in future work.

5. CONCLUSION

As a potential tool for psychotherapy, the presented work focused on the tasks of sentiment detection and the classification of self-assessed emotion labels in personal narratives. It includes the first attempt at modelling shifts in state-of-mind (SoM) from linguistic data as reflected in changes in self-assessed arousal and valence scores. We proposed and developed hierarchical attention models which operate on a complex dataset with long spoken narratives and weak labels. Our approach not only achieved near state-of-the-art results when compared to acoustic analysis, but also extracted interesting word uses related to the sentiment of each narrative.

One limitation of our approach is that 10-class classification using the entire range of the self-assessed affect scores was not overly successful. We speculate that this was due to inconsistencies inherent in self-assessed labels and the relatively small dataset. We aim to address this in future work by exploring transfer learning approaches which can learn a joint representation through shared weights. Other work will include repeating our analysis on similar corpora for tasks such as depression and narcissism detection.

6. ACKNOWLEDGEMENTS

This research has received funding from BMW Group Research and the EU's 7$^{\text{th}}$ Framework Programme ERC Starting Grant No. 338164 (iHEARu).

7. REFERENCES

[1] J.A. Russell, "Core affect and the psychological construction of emotion," *Psychological review*, vol. 110, no. 1, pp. 145–172, 2003.

[2] N. Schwarz and G. L. Clore, "Mood, misattribution, and judgments of well-being: informative and directive functions of affective states.," *Journal of personality and social psychology*, vol. 45, no. 3, pp. 513–523, 1983.

[3] K. Grawe, *Neuropsychotherapie*, Hogrefe Verlag, Göttingen, Germany, 2004.

[4] E.-M. Rathner, Y. Terhorst, N. Cummins, B. Schuller, and H. Baumeister, "State of Mind: Classification through Self-reported Affect and Word Use in Speech.," in *Proceedings INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, September 2018, pp. 267–271, ISCA.

[5] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A Survey of Multimodal Sentiment Analysis," *Image and Vision Computing, Special Issue on Multimodal Sentiment Analysis and Mining in the Wild*, vol. 35, pp. 3–14, 2017.

[6] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, Seattle, WA, 2013, pp. 1631–1642, ACL.

[7] O. Irsoy and C. Cardie, "Opinion mining with deep recurrent neural networks," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar, 2014, pp. 720–728, ACL.

[8] D. Tang, F. Wei, B. Qin, T. Liu, and M. Zhou, "Coooolll: A deep learning system for twitter sentiment classification," in *Proceedings of the 8th International Workshop on Semantic Evaluation*, Dublin, Ireland, 2014, pp. 208–212, ACL.

[9] B. Schuller, S. Steidl, A. Batliner, P. B. Marschik, H. Baumeister, F. Dong, S. Hantke, F. B. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 Computational Paralinguistics Challenge: Atypical and Self-Assessed Affect, Crying and Heart Beats," in *Proceedings INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, September 2018, pp. 122–126, ISCA.

[10] 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*. 2016, pp. 1480–1489, ACL.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., pp. 5998–6008. Curran Associates, Inc., 2017.

[12] J. Cheng, L. Dong, and M. Lapata, "Long short-term memory-networks for machine reading," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, TX, 2016, pp. 551–561, ACL.

[13] A. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, TX, 2016, pp. 2249–2255, ACL.

[14] H. Chen, M. Sun, C. Tu, Y. Lin, and Z. Liu, "Neural sentiment classification with user and product attention," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, TX, 2016, pp. 1650–1659, ACL.

[15] C. Kwak, J. Son, A. Lee, and S. Kim, "Scene emotion detection using closed caption based on hierarchical attention network," in *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju, South Korea, 2017, pp. 1206–1208, IEEE.

[16] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis," *IEEE Intelligent Systems Magazine*, vol. 28, no. 2, pp. 15–21, March/April 2013.

[17] B. Schuller, A. E.-D. Mousa, and V. Vasileios, "Sentiment Analysis and Opinion Mining: On Optimal Parameters and Performances," *WIREs Data Mining and Knowledge Discovery*, vol. 5, pp. 255–263, September/October 2015.

[18] B. Schuller, "Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, May 2018.

[19] F. Weninger, M. Wöllmer, and B. Schuller, "Emotion Recognition in Naturalistic Speech and Language – A Survey," in *Emotion Recognition: A Pattern Analysis Approach*, A. Konar and A. Chakraborty, Eds., chapter 10, pp. 237–267. Wiley, 1st edition, December 2015.

[20] C. Gorrostieta, R. Brutti, K. Taylor, A. Shapiro, J. Moran, A. Azarbayejani, and J. Kane, "Attention-based sequence classification for affect detection," in *Proceedings INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, September 2018, pp. 506–510, ISCA.

[21] C. Montacié and M.-J. Caraty, "Vocalic, lexical and prosodic cues for the interspeech 2018 self-assessed affect challenge," in *Proceedings INTERSPEECH 2018, 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, September 2018, pp. 541–545, ISCA.

[22] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "AVEC 2017 – Real-life Depression, and Affect Recognition Workshop and Challenge," in *Proceedings of the 7th International Workshop on Audio/Visual Emotion Challenge, AVEC'17, co-located with the 25th ACM International Conference on Multimedia, MM 2017*, Mountain View, CA, October 2017, pp. 3–9, ACM.

[23] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," `https://arxiv.org/abs/1406.1078`, 2014.

[24] J. Kimball, "Seven principles of surface structure parsing in natural language," *Cognition*, vol. 2, no. 1, pp. 15–47, 1973.

[25] W. Ammar, G. Mulcaire, Y. Tsvetkov, G. Lample, C. Dyer, and N. A. Smith, "Massively multilingual word embeddings," `https://arxiv.org/abs/1602.01925`, 2016.