KERMIT for Sentiment Analysis in Italian Healthcare Reviews

Leonardo Ranaldi[1], **Michele Mastromattei**[2], **Dario Onorati**[2],
Elena Sofia Ruzzetti[2], **Francesca Fallucchi**[1], **Fabio Massimo Zanzotto**[2]
1. Dept. of Innovation and Information Engineering Guglielmo Marconi University, Italy
2. Dept. of Enterprise Engineering University of Rome Tor Vergata, Italy
l.ranaldi@unimarconi.com, elenasofia.ruzzetti@alumni.uniroma2.eu,
{michele.mastromattei,fabio.massimo.zanzotto}@uniroma2.it,
dario.onorati@uniroma1.it, f.fallucchi@unimarconi.it

Abstract

English. In this paper, we describe our approach to the sentiment classification challenge on Italian reviews in the healthcare domain. Firstly, we followed the work of Bacco et al. (2020) from which we obtained the dataset. Then, we generated our model called KERMIT$_{HC}$ based on *KERMIT* (Zanzotto et al., 2020). Through an extensive comparative analysis of the results obtained, we showed how the use of syntax can improve performance in terms of both accuracy and F1-score compared to previously proposed models. Finally, we explored the interpretative power of *KERMIT*-viz to explain the inferences made by neural networks on examples.

Italiano. *In questo lavoro, presentiamo il nostro approccio al task di sentiment analysis per le recensioni italiane in ambito sanitario. Abbiamo seguito il lavoro di Bacco et al. (2020) da cui abbiamo ottenuto il dataset. Successivamente, abbiamo usato KERMIT$_{HC}$ basato su* KERMIT*(Zanzotto et al., 2020). Da un'ampia analisi comparativa dei risultati ottenuti mostriamo come l'uso della sintassi può migliorare le prestazioni sia in termini di accuratezza che di F1-score rispetto ai modelli proposti in precedenza. Infine, abbiamo esplorato il potere interpretativo di* KERMIT*-viz per spiegare le inferenze fatte dalle reti neurali sugli esempi.*

1 Introduction

People are practically reviewing anything in online sites and understanding the polarization of a comment through automatic sentiment classifier is a tantalizing challenge. In recent years, the number of virtual reviewers has drastically increased and there are many products and services, which can be reviewed. Each person, before buying a product or a service, searches into reviews from people who have already had experienced the product or the service. Review portals are usually linked to the leisure or business activities such as the world of tourism, e-commerce or movies. However, there are topics where these reviews and the associated automatic computed sentiment may induce to select wrong services, which may dramatically affect personal life.

When dealing with health-related services, the effect of positive or negative reviews on hospitals and doctors can have a potential catastrophic impact on the health of who is using this piece of information. QSalute [1] is one of the most important Italian portals of reviews about hospitals, nursing homes and doctors. It is very important for patients to seek the best hospital for their condition based on the past experience of other patients. Reviews in the world of health benefit both patients and hospitals because they are a means to discover problems and solve them (Greaves et al., 2013; Khanbhai et al., 2021).

Automatic sentiment analyzer have then a big responsibility in the context of health-related services. In these sensitive areas, it is important to design AI systems whose decisions are transparent (Doshi-Velez and Kim, 2017), that is, the systems must give the motivation for the choice made so that people can trust. If the users do not trust a

[1] https://www.qsalute.it/

model or a prediction, they will not use it (Ribeiro et al., 2016).

In this article, we investigate a model that can mitigate the responsibility of sentiment analyzers for health-related services. The model we are using exploits syntactic information within neural networks to provide a clear visualisation of the internal decision mechanism of the model that produced the decision. We propose KERMIT$_{HC}$ (*KERMIT* for **H**ealth**C**are) based on *KERMIT* (Zanzotto et al., 2020) to solve the sentiment analysis task introduced by Bacco el al.(2020). We use KERMIT$_{HC}$ on QSalute Italian portal reviews in order to include symbolic knowledge as a part of the architecture and visualize the internal decision-making mechanism of the neural model, using *KERMIT*-viz (Ranaldi et al., 2021).

In the rest of paper, Section 2 gives details about the dataset and methods, while Section 3 and 4 describe the experiments, the results obtained and their discussion. Finally, in Section 5 we present the final conclusions and future goals.

2 Data & Methods

To explore our hunch that syntactic interpretation may help in Healthcare reviews recognition, we leverage: (1) a Healthcare training corpus (Sec. 2.1); (2) a KERMIT$_{HC}$, which is based on syntactic interpretation and it can explain its decisions; and finally, (3) some challenges solved due to KERMIT$_{HC}$ (Sec. 2.2).

2.1 Dataset

In order to investigate reviews in healthcare area, we selected the QSalute portal, one of the most important health websites in Italy. This portal can be defined as the TripAdvisor of hospital facilities, indeed it talks about: *Expertise*, *Assistance*, *Cleaning* and *Services*. In addition to the reviews, there are some associated metadata such as: *user id*, *hospital name*, *review title* and *patient pathology*. To ensure privacy we do not consider sensitive data such as *user id* and *hospital name*.

We used a free available scraper on GitHub [2] to download the dataset. Then, to model this data to a sentiment analysis task, we followed the indications provided by Bacco et al.(2020) - in detail, a review is: (1) negative if the average of its scores is less than or equal to 2, (2) positive if the average of its scores is greater than or equal to 4 (3) neutral otherwise.

The resulting dataset is composed of 47,224 reviews consisting of: 40,641 reviews in the positive class, 3,898 in the neutral class and 2,685 in the negative class.

In this work, we solely consider positive and negative classes, so our final dataset is composed of 43,326 reviews. The dataset is heavily skewed (93,80% positive class - 6,20% negative class) favoring reviews labeled as positive.

2.2 KERMIT 4 Healthcare

KERMIT$_{HC}$ (*KERMIT* for **H**ealth**C**are) architecture is composed of 3 major parts: (1) a *KERMIT* model described in Zanzotto et al. (2020), (2) a Transformers model and (3) a decoder layer that combines the results obtained from the previous two sub-parts. In figure Fig.1 we show a graphical representation of the architecture of KERMIT$_{HC}$, pointing the parts that compose it.

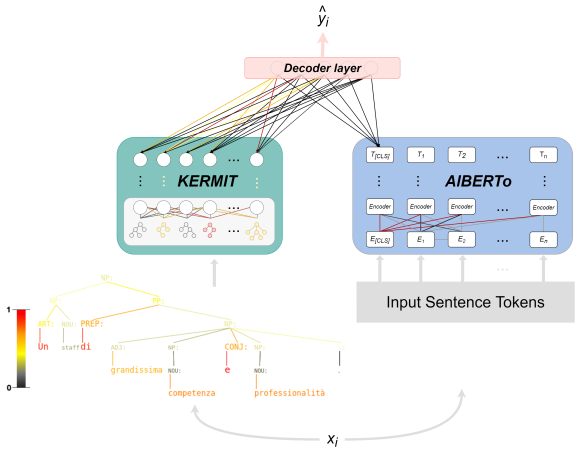


Figure 1: KERMIT$_{HC}$ architecture, forward and interpretation pass.

The architecture of KERMIT$_{HC}$ makes it a particular model, because it combines the syntax offered by KERMIT with the versatility of a Transformer-model. We use KERMIT because it allows the encoding of universal syntactic interpretations in a neural network architecture. KERMIT component is itself composed of two parts: KERMIT encoder, which converts parse tree T into embedding vectors and a multi-layer perceptron that exploits these embedding vectors. The second sub-part of our architecture is composed of a Bidirectional Encoder Representations from Transformers, - as known as BERT - to classify the

[2] The scraper is available at `https://github.com/lbacco/Italian-Healthcare-Reviews-4-Sentiment-Analysis`

Model	Average Accuracy	Average Macro F1 score	Average Weighed F1 score
UmBERTo	$0.74(\pm 0.14)^{\diamond}$	$0.43(\pm 0.02)$	$0.75(\pm 0.18)^{\circ}$
AlBERTo	$\mathbf{0.82(\pm 0.15)}^{\diamond}$	$\mathbf{0.47(\pm 0.05)}^{\dagger}$	$\mathbf{0.8(\pm 0.14)}^{\circ}$
BERT multilingual	$0.73(\pm 0.13)$	$0.46(\pm 0.1)^{\dagger}$	$0.73(\pm 0.22)$
ELECTRA$_{ita}$	$0.67(\pm 0.17)$	$0.4(\pm 0.13)$	$0.66(\pm 0.2)$

Table 1: Performance of *BERT*, on 25% of the QSalute dataset. Mean and standard deviation results are obtained from 10 runs. For each *Site*, the best performing model was highlighted based on the F1 score values obtained. The symbols $\diamond$, $\circ$ and † indicate a statistically significant difference between two results with a 95% of confidence level with the sign test.

sentiment of the reviews. BERT is a pre-trained language model developed by Devlin et al. (2019) at Google AI Language. In particular, since the task concerns sentences in the Italian language, we have used a special BERT version pretrained on that language called AlBERTo (Polignano et al., 2019).

3 Experiments

We used KERMIT$_{HC}$ architecture to examine if it is possible to answer the research questions showed in *KERMIT* (Zanzotto et al., 2020) also in healthcare domain using the Italian language. Those research questions are: (1) Can the *symbolic knowledge* provided by universal symbolic syntactic interpretations, make a difference and it be used effectively in neural networks? (2) Do *universal symbolic syntactic interpretations* encode different syntactic information than those encoded in *"embeddings of universal sentences"*? (3) Can the *universal symbolic syntactic interpretations* provided by KERMIT$_{HC}$, supply a better and clearer way to explain the decisions of neural networks than those provided by transformers?

To provide a comprehensive answer to these questions, we tested the architecture in a *completely universal* setting where both *KERMIT* and AlBERTo are trained only in the last decision layer.

The rest of the Section describes the experimental set-up, the quantitative experimental results and discusses how we can use the *KERMIT-viz* to explain decisions of neural network inferences over examples.

3.1 Experimental Set-up

This section describes the general experimental set-up of our experiments and the specific configurations adopted.

The parameters used for the *KERMIT* encoder are those proposed in Zanzotto et al., (2020) paper. The constituency parse trees used for *KERMIT* sub-part are obtained using our freely available script on GitHub[3].

We tested several different BERT version pretrained on Italian language in order to get the best model for our task. In particular, we tested the following transformers: (1) UmBERTo (Parisi et al., 2020); (2) AlBERTo (Polignano et al., 2019); (3) BERT multilingual (Devlin et al., 2018) and (4) ELECTRA$_{ita}$: an Italian version of ELECTRA model (Clark et al., 2020) implemented by Schweter (2020) on a work of Chan et al. (2020). All the models were implemented using Huggingface's transformers library (Wolf et al., 2019) and all were used in the uncased setting with the pretrained version. The input text for BERT has been preprocessed and tokenized as specified in respectively work (Parisi et al., 2020; Polignano et al., 2019; Devlin et al., 2018; Schweter, 2020).

Since our experiments are text classification task, the decoder layer of our KERMIT$_{HC}$ architecture is a fully connected layer with the softmax activation function applied to the concatenation of the *KERMIT* sub-part output and the final `[CLS]` token representation of the selected transformer model. Finally, the optimizer used to train the whole architecture is AdamW (Loshchilov and Hutter, 2019) with the learning rate set to $2e^{-5}$. For reproducibility, the source code of our experiments is publicly available on our GitHub repository[4].

[3]The code is available at `https://github.com/LeonardRanaldi/Constituency-Parser-Italian`

[4]The code is available at `https://github.com/ART-Group-it/KERMIT-4-Sentiment-Analysis-on-Italian-Reviews-in-Healthcare`

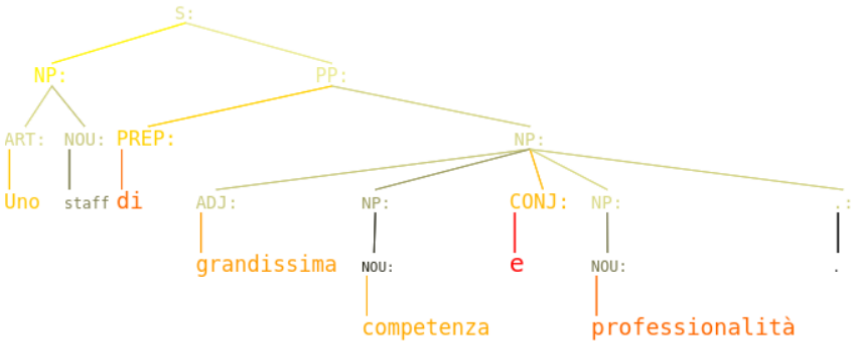
Site	Model	Average Accuracy	Average Macro F1 score	Average Weighed F1 score
Pneumology	**KERMIT$_{HC}$**	**0.71 (± 0.14)**	**0.51 (± 0.08)**	**0.7 (± 0.11)**
	AlBERTo	0.66 (± 0.27)	0.4 (± 0.12)†	0.61 (± 0.26)
Thoracic Surgery	**KERMIT$_{HC}$**	**0.78 (± 0.13)**	**0.51 (± 0.07)**	**0.81 (± 0.08)**
	AlBERTo	0.74 (± 0.28)	0.43 (± 0.13)	0.74 (± 0.26)
Nervous System	**KERMIT$_{HC}$**	**0.87 (± 0.05)**†	**0.6 (± 0.03)**†	**0.89 (± 0.03)**
	AlBERTo	0.94 (± 0.01)†	0.48 (± 0.0)†	0.91 (± 0.01)
Hearth	**KERMIT$_{HC}$**	**0.93 (± 0.03)**†	**0.56 (± 0.03)**†	**0.93 (± 0.02)**
	AlBERTo	0.96 (± 0.01)†	0.49 (± 0.0)†	0.94 (± 0.01)
Vascular Surgery	**KERMIT$_{HC}$**	**0.81 (± 0.16)**	**0.49 (± 0.06)**†	**0.83 (± 0.12)**
	AlBERTo	0.70 (± 0.29)	0.42 (± 0.11)†	0.73 (± 0.23)
Ophthalmology	**KERMIT$_{HC}$**	**0.79 (± 0.08)**	**0.55 (± 0.05)**†	**0.83 (± 0.06)**
	AlBERTo	0.87 (± 0.08)	0.48 (± 0.02)†	0.86 (± 0.04)
Rheumatology	**KERMIT$_{HC}$**	**0.58 (± 0.23)**	**0.43 (± 0.11)**	**0.60 (± 0.20)**
	AlBERTo	0.68 (± 0.20)	0.44 (± 0.10)	0.69 (± 0.19)
Infections	**KERMIT$_{HC}$**	**0.68 (± 0.19)**	**0.51 (± 0.12)**	**0.70 (± 0.17)**
	AlBERTo	0.57 (± 0.23)	0.42 (± 0.13)	0.58 (± 0.21)
Skin	**KERMIT$_{HC}$**	**0.64 (± 0.11)**	**0.50 (± 0.07)**	**0.70 (± 0.10)**
	AlBERTo	0.63 (± 0.26)	0.39 (± 0.11)	0.61 (± 0.24)
Genital	**KERMIT$_{HC}$**	**0.79 (± 0.09)**†	**0.55 (± 0.03)**†	**0.82 (± 0.06)**
	AlBERTo	0.88 (± 0.06)†	0.49 (± 0.02)†	0.87 (± 0.03)
Endoscopy	**KERMIT$_{HC}$**	**0.75 (± 0.09)**	**0.52 (± 0.04)**†	**0.80 (± 0.05)**
	AlBERTo	0.80 (± 0.19)	0.45 (± 0.07)†	0.78 (± 0.17)
Facial	**KERMIT$_{HC}$**	**0.70 (± 0.24)**	**0.42 (± 0.08)**	**0.76 (± 0.18)**
	AlBERTo	0.72 (± 0.26)	0.42 (± 0.10)	0.76 (± 0.22)
Oncology	**KERMIT$_{HC}$**	**0.91 (± 0.06)**	**0.52 (± 0.04)**†	**0.92 (± 0.03)**
	AlBERTo	0.89 (± 0.21)	0.46 (± 0.08)†	0.89 (± 0.17)
Haematology	**KERMIT$_{HC}$**	**0.56 (± 0.30)**	**0.36 (± 0.14)**	**0.57 (± 0.31)**
	AlBERTo	0.41 (± 0.25)	0.30 (± 0.11)	0.46 (± 0.23)
Endocrinology	**KERMIT$_{HC}$**	**0.71 (± 0.20)**	**0.48 (± 0.12)**	**0.71 (± 0.22)**
	AlBERTo	0.73 (± 0.29)	0.41 (± 0.13)	0.69 (± 0.28)
Gynaecology	**KERMIT$_{HC}$**	**0.82 (± 0.08)**	**0.56 (± 0.05)**†	**0.85 (± 0.05)**
	AlBERTo	0.85 (± 0.14)	0.48 (± 0.04)†	0.84 (± 0.09)
Otorhinology	**KERMIT$_{HC}$**	**0.84 (± 0.14)**	**0.50 (± 0.06)**	**0.86 (± 0.09)**
	AlBERTo	0.80 (± 0.18)	0.46 (± 0.05)	0.83 (± 0.13)

Table 2: Performance of KERMIT$_{HC}$ and AlBERTo on QSalute database grouped by *Site*. Mean and standard deviation results are obtained from 10 runs. For each *Site*, the best performing model was highlighted based on the F1 score values obtained. The symbol † indicate a statistically significant difference between two results with a 95% of confidence level with the sign test.

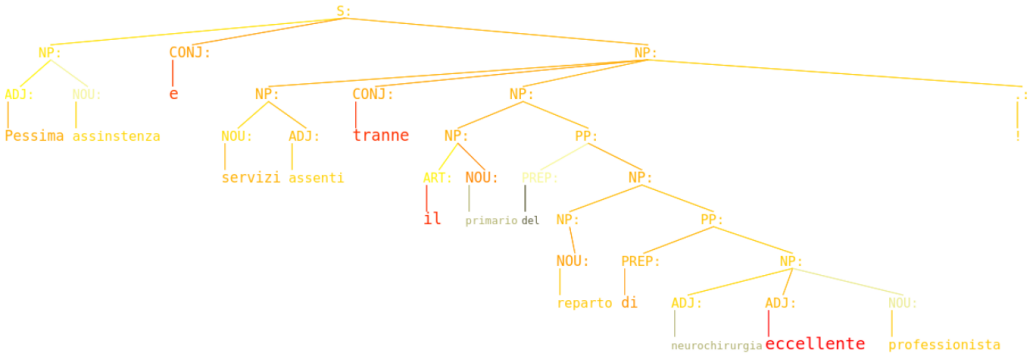
4 Results and Discussion

Syntactic information is useful to significantly increase performances to classify Healthcare reviews (see Table 2). KERMIT$_{HC}$ uses AlBERTo which is the best BERT-italian version model according to our experiments, showed in Table 1. Especially KERMIT$_{HC}$ outperforms the solely AlBERTo sub-part model (ref. to Table 2).

As in the work proposed by Bacco et al.(2020), we chose to divide the dataset by *"Site"* and evaluate the models using accuracy and F1-score metrics. Despite this division, the dataset is still very unbalanced favoring the class 1 (positive reviews). We reports results in terms of the accuracy, Macro F1 and Weighed F1. Observing Table 2, we can see that the performance obtained by KERMIT$_{HC}$ always exceeds the best configuration of *BERT*: AlBERTo. Hence, trained on the Healthcare review dataset (Bacco et al., 2020) (see Section 2.1) KERMIT$_{HC}$ seems to be a good candidate to analyze sentiment of hospital patients.



(a) **S:** `Uno staff di grandissima competenza e professionalità!`



(b) **S:**`Pessima assistenza e servizi assenti tranne il primario di reparto di neurochirurgia eccellente professionista`

Figure 2: The visualizations offered by *KERMIT-viz*. Both examples have the target class positive but in the first one, it is easy to state the positivity. In the second one, who wrote the review, makes disquisitions about the medical staff but at the same time lauds the head of the department.

Using the *KERMIT-viz* visualiser, we analysed how important the contribution of symbolic knowledge provided by *KERMIT* can be. In many cases it makes all the difference. Looking at the Figure 2, these are two sentences with a positive target. The first sentence (shown in Fig. 2a) is clearly positive while the sentence shown in the Fig. 2b could be ambiguous as the patient makes bad remarks about the service but praises the head of the department. We can observe how some words have been colored in red (therefore they have received a greater weight during the classification phase) emphasizing the positive aspects of the sentence and causing it to be labeled as *"positive review"*. In this way the explainability is guaranteed and in very delicate topics - like sentiment in health reviews - we can have more *"trust"* on sentiment analysers.

5 Conclusion

In this article, we investigated a model that can mitigate the responsibility of sentiment analyzers for health-related services. Our model KERMIT$_{HC}$ exploits syntactic information within neural networks to provide a clear visualisation of its internal decision mechanism. KERMIT$_{HC}$ is based on *KERMIT* (Zanzotto et al., 2020) and we worked in a sentiment analysis task introduced by Bacco el al.(2020).

We studied several versions of pre-trained BERT models on the Italian language and found out that AlBERTo is, among them, the best model for this task. However, KERMIT$_{HC}$, which is composed of *KERMIT*+AlBERTo, outperforms better than AlBERTo model alone. Additionally, via *KERMIT-viz*, we visualized the reasons why KERMIT$_{HC}$ classifies the dataset. We observed how KERMIT$_{HC}$ captures relevant syntactic information by catching the keywords in each sen-

tence giving them more weight in the decision phase, mitigating and capturing possible errors of the sentiment analysers. Our future goal is to be able to have full control of the sentiment analysers by injecting human rules (Onorati et al., 2020) in order to mitigate possible errors.

References

Luca Bacco, A. Cimino, L. Paulon, M. Merone, and F. Dell'Orletta. 2020. A machine learning approach for sentiment analysis for italian reviews in healthcare. In *CLiC-it*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning.

Felix Greaves, Daniel Ramirez-Cano, Christopher Millett, Ara Darzi, and Liam Donaldson. 2013. Use of sentiment analysis for capturing patient experience from free-text comments posted online. *Journal of medical Internet research*, 15:e239, 11.

Mustafa Khanbhai, Patrick Anyadi, Joshua Symons, Kelsey Flott, Ara Darzi, and Erik Mayer. 2021. Applying natural language processing and machine learning techniques to patient experience feedback: a systematic review. *BMJ Health & Care Informatics*, 28(1).

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *7th International Conference on Learning Representations, ICLR 2019*.

Dario Onorati, Pierfrancesco Tommasino, Leonardo Ranaldi, Francesca Fallucchi, and Fabio Massimo Zanzotto. 2020. Pat-in-the-loop: Declarative knowledge for controlling neural networks. *Future Internet*, 12(12).

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: An italian language model trained with whole word masking.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, Giovanni Semeraro, and Valerio Basile. 2019. AlBERTo: Italian BERT Language Understanding Model for NLP Challenging Tasks Based on Tweets. In *Proceedings of the Sixth Italian Conference on Computational Linguistics (CLiC-it 2019)*, volume 2481. CEUR.

Leonardo Ranaldi, Francesca Fallucchi, and Fabio Massimo Zanzotto. 2021. KERMITviz: Visualizing Neural Network Activations on Syntactic Trees. In *In the 15th International Conference on Metadata and Semantics Research (MTSR'21)*, volume 1.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier.

Stefan Schweter. 2020. Italian bert and electra models, November.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.0.

Fabio Massimo Zanzotto, Andrea Santilli, Leonardo Ranaldi, Dario Onorati, Pierfrancesco Tommasino, and Francesca Fallucchi. 2020. KERMIT: Complementing transformer architectures with encoders of explicit syntactic interpretations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 256–267, Online, November. Association for Computational Linguistics.