Human Perception in Natural Language Generation

de Mattei, Lorenzo; Lai, Huiyuan; Dell'Orletta, Felice; Nissim, Malvina

Published in:
Proceedings of the 1st Workshop on Generation Evaluation and Metrics

Publication date:
2021

Human Perception in Natural Language Generation

Lorenzo De Mattei[⋆◇♣‡], **Huiyuan Lai**[♣], **Felice Dell'Orletta**[◇], **Malvina Nissim**[♣]
[⋆] Department of Computer Science, University of Pisa / Italy
[◇] ItaliaNLP Lab, Istituto di Linguistica Computazionale "Antonio Zampolli", Pisa / Italy
[♣] CLCG, University of Groningen / The Netherlands
[‡] Aptus.AI / Pisa, Italy
lorenzo.dematteі@di.unipi.it
{h.lai,m.nissim}@rug.nl
felice.dellorletta@ilc.cnr.it

Abstract

We take a collection of short texts, some of which are human-written, while others are automatically generated, and ask subjects, who are unaware of the texts' source, whether they perceive them as human-produced. We use this data to fine-tune a GPT-2 model to push it to generate more human-like texts, and observe that the production of this fine-tuned model is indeed perceived as more human-like than that of the original model. Contextually, we show that our automatic evaluation strategy correlates well with human judgements. We also run a linguistic analysis to unveil the characteristics of human- vs machine-perceived language.

1 Introduction

Pre-trained language models, such as the BERT (Devlin et al., 2019) and the GPT (Radford et al., 2018, 2019) families, are nowadays the core component of NLP systems. These models, based on the Transformer (Vaswani et al., 2017) and trained using huge amounts of crawl data (which can contain substantial noise), have been shown to produce high quality text, more often than not judged as human-written (Radford et al., 2019; De Mattei et al., 2020; Brown et al., 2020). Existing evaluations of GPT-2 models (Ippolito et al., 2020; De Mattei et al., 2020) have shown that while generated sentences were ranked lower in human perception than gold sentences, many gold sentences were also not perceived as human-like. To make the model produce more human-like texts one could train it only on gold data which is highly perceived as human, but such data is costly, and full model retraining is often a computationally non-viable option. As an alternative route, we explore whether and how an existing pre-trained model can be instead *fine-tuned* to produce more humanly-perceived texts, and how to evaluate this potentially shifted behaviour.

We see the advantage of this experiment at least in two ways. One is that the generation of more human-like texts is highly beneficial for specific applications, as for example human-machine interaction in dialogues; the other is that it opens the opportunity to investigate what linguistic aspects make a text more humanly-perceived. We run our experiments on Italian, using `GePpeTto` (De Mattei et al., 2020) as pre-trained model. First, we collect human judgements on gold texts and texts generated by `GePpeTto` in terms of how they are perceived (human or automatically produced). We then fine-tune `GePpeTto` with this perception-labelled data. In addition, inspired by the classifier-based reward used in style transfer tasks (Lample et al., 2019; Gong et al., 2019; Luo et al., 2019; Sancheti et al., 2020), we reward the model to push its classification confidence. We evaluate the new perception-enhanced models in comparison with the original `GePpeTto` by running both an automatic as well as a human evaluation on output generated by the various models. Lastly, we conduct a linguistic analysis to highlight which linguistic characteristics are more commonly found in human- and machine-perceived text.

Author contribution note: Lorenzo De Mattei and Huiyuan Lai contributed equally.

Contributions We show that a GPT-2 pre-trained model can be fine-tuned to produce text that is perceived as more human, and we release this model for Italian. Second, we provide a stronger automatic evaluation method where training is done on perception labels rather than the actual source, which yields results that correlate with human judgments, providing a different angle for automatic evaluation of generated sentences. Lastly, we run a linguistic analysis of the humanly-perceived texts that can open up to new opportunities for understanding and model human-like perception.

2 Data

We collected human judgments over a series of gold and generated sentences in terms of how much a given text is *perceived* as human-like. The obtained labelled data is used to fine-tune our base model towards generating more humanly-perceived texts; it is also used to test the resulting models through an automatic evaluation strategy that we implement next to human judgements.

Training Data From the original `GePpeTto`'s training corpus (De Mattei et al., 2020), we collected 1400 random gold sentences in the following way. We sentence split all the documents and we picked the first sentence of each document. In order to allow for length variation, which has an impact on perception, we selected the first 200 sentences with length 10, 15, 20, 25, 30, 35 and 40 tokens.

We also let `GePpeTto` generate texts starting with the first word of randomly selected documents, we sentence-split the generated texts, and select the first 200 sentences with length 10, 15, 20, 25, 30, 35 and 40 tokens. This procedure creates a training set with perception labels containing a total of 2800 instances (1400 gold and 1400 generated).

We asked native Italian speakers if they felt the text they were seeing had been written, on a 1–5 Likert Scale, by a human (1) or a machine (5). Each texts was assessed by 7 different judges. The subjects for the task were laypeople recruited via the crowdsourcing platform Prolific[1]. We did not control for, and thus did not elicit, any demographic features. As a proxy for attention and quality control, we used completion time, and filtered out participants who took too little time to perform the task (we set a threshold of at least 5 minutes for 70 assessments as a reliable minimum effort).[2]

Mapping the average of human judgements to a binary classification (human if < 3), we obtain the matrix in Tab. 1 showing perception labels and the actual source labels. While human texts are more often perceived as human-like than machine-generated ones, the matrix shows that 44.2% of the texts are perceived as artificial, suggesting that a good portion of the training data might lead to generation that is not so much human-like. We train two classifiers on 80% of this data on the task of detecting human-like perception and that of detecting the actual source. The classifiers are built adding a dropout (Srivastava et al., 2014) and a dense layer on the top of UmBERTo[3], which is a Roberta (Liu et al., 2019) based Language Model trained on large Italian corpora. We train them using Adam (Kingma and Ba, 2015), initial learning rate 1e-5, and batch size 16. On the remaining 20% of the data we obtain F=0.97 for the source identification task, and F=0.92 for the perception task, showing the feasibility of the classification and thus the possibility of using these classifiers for evaluation (Section 4).

	AI-perceived	humanly-perceived
`GePpeTto`	62.3%	37.7%
Gold	44.2%	55.8%

Table 1: Source vs perception matrix (training data).

Test Data We use 1400 sentences: 350 are produced by humans, 1050 are generated (350 for each of the three models we use, see Section 3). As for training, human texts were selected picking the first 50 sentences with 10, 15, 20, 25, 30, 35 and 40 tokens. For each system, we also picked the first 50 generated sentences with length 10, 15, 20, 25, 30, 35 and 40 tokens. Each of the 1400 sentences was assessed by 5 users, on a 1–5 Likert scale, as human- or artificial-like.

3 Models

We use three models for text generation, all based on the GPT-2 architecture (Radford et al., 2019). The basic model is `GePpeTto`, a GPT-2-based model for Italian released by (De Mattei et al., 2020). The others are built on `GePpeTto` using

[1] https://www.prolific.co/

[2] Crowdworkers were compensated with a rate of £5.04 per estimated hour. In practice, tasks were completed in a shorter time than estimated, so the hourly rate was a bit higher.

[3] `https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1`

the perception-labelled data in fine-tuning and in a reinforcement learning setting.

3.1 GePpeTto

GePpeTto is built using GPT-2 base architecture with 12 layers and 117M parameters. GePpeTto is trained on two main sources: a dump of Italian Wikipedia, consisting of 2.8GB of text; and the ItWac corpus (Baroni et al., 2009), which amounts to 11GB of web texts. De Mattei et al. (2020) show that GePpeTto is able to produce text which is much closer to human quality rather than to the text generated by other baseline models. Still, real human-produced text is recognised as such more often than GePpeTto's output.

3.2 GePpeTto fine-tuned

Using the original settings of GePpeTto, the model is fine-tuned on the training portion of the humanly-perceived sentences of the perception-labelled data (Tab. 1), using the Huggingface implementation (Wolf et al., 2020).[4]. We use the Adam optimiser (Kingma and Ba, 2015) with initial learning rate 2e-5. The mini-batch size is set to 8. During fine-tuning, we set an early stopping with patience 5 if the performance on validation does not improve.[5] The resulting model should produce text recognised more frequently as human-produced than the original GePpeTto.

3.3 GePpeTto rewarded

To further encourage GePpeTto-F to generate more humanly-perceived texts, we introduce a confidence reward based on the 'perception classifier' (PC) described in Section 2: the model gets rewarded for generating more human-like text. The PC's confidence is formulated as

$$R_{conf} = softmax_0(PC(\boldsymbol{y}', \theta)) \quad (1)$$

where θ are the PC's parameters, fixed during fine-tuning GePpeTto . Formally, the confidence is used for policy learning that maximizes the expected reward $E[R]$ of the generated sequence; the corresponding policy gradient is formulated as

$$\nabla_\phi E(R) = \nabla_\phi \sum_k (P(\boldsymbol{y}_t^s | \boldsymbol{y}_{1:t-1}^s; \phi))R_k \quad (2)$$

where ϕ are the parameters of GePpeTto, and R_k is the reward of the k_{th} sequence $\boldsymbol{y}^s$ sampled from the distribution of model's outputs at each time step in decoding. The framework can be trained end-to-end by combining the policy gradient with the cross entropy loss of the base model.

4 Evaluation

We run both a human and an automatic evaluation, in line with Ippolito et al. (2020)'s and Hashimoto et al. (2019)'s suggestions in terms of evaluation's diversity and quality. For the automatic evaluation, we train a regressor on the perception-labelled data (with the original 1–5 values) adding a dropout (Srivastava et al., 2014) and a dense layer on the top of UmBERTo. We use Adam (Kingma and Ba, 2015) with initial learning rate is 1e-5, and set the batch size to 16. We calculate the correlation of the regressor's scores with human judgements over each single data point in the test set (N=1400), and observe good scores (Pearson=0.54 ($p < 10^{-4}$) and RMSE=0.75).

For the human evaluation, we assign to each sentence the average score computed over all human judgements. We then average all resulting scores over the seven length bins. Results are shown in two tables, as follows.

First, as we did for the training data (see Table 1), we mapped the average of human judgements to a binary classification (human if< 3), and obtain the matrix in Table 2. This shows perception labels and the actual source labels for the three models and gold data. We see that the human produced texts are the most humanly-perceived, but both the fine-tuned and the rewarded model produced texts that are more humanly-perceived than GePpeTto, with the fine-tuned model performing better than the rewarded one.

Second, Table 3 shows the average score over all length bins for the four models: GePpeTto, GePpeTto fine-tuned (GePpeTto-F), GePpeTto rewarded (GePpeTto-R) and the original human texts (Human). This table also reports the average scores over all lengths as assigned by the regressor.[6] The closer to 1, the more humanly-

[4] In preliminary experiments, we also fine-tuned GePpeTto on a larger silver data-set obtained by letting the perception classifier select what it deemed are human-perceived texts from GePpeTto's training set. The results of our automatic evaluation were however not encouraging, suggesting that the increased performance we obtain with the fine-tuned model is indeed ascribable to manually labelled gold data.

[5] Due to small training size, we validate against silver data obtained by labelling generated and gold text with our perception-classifier.

[6] Detailed results per length are Appendix Tables A.1-A.2.

perceived the sentence.

	AI-perceived	humanly-perceived
GePpeTto	61.1%	38.96%
GePpeTto-F	55.7%	44.3%
GePpeTto-R	59.1%	40.9%
Gold	37.4%	62.6%

Table 2: Source vs perception matrix (test data).

model	humans (std)	regressor (std)
GePpeTto	2.85 (0.83)	2.74 (0.71)
GePpeTto-F	2.74 (0.83)	2.49 (0.55)
GePpeTto-R	2.84 (0.87)	2.56 (0.57)
Human	2.41 (0.77)	2.47 (0.66)
avg	2.71 (0.85)	2.57 (0.63)

Table 3: Scores for each system as evaluated by humans and by the regressor, averaged over test set instances and thus over all sentence lengths.

As a first observation, in both the human and the automatic evaluations the final rank for the systems is the same, showing the reliability of the automatic evaluation. The gold texts are perceived as most human-like by humans (score: 2.41) and by the regressor (score: 2.47). Regarding systems, the fine-tuned model (GePpeTto-F) performs better than both the basic and the rewarded model.

To compare the overall performance of machine vs humans, in Fig 1 we plot the average performance of the three models per length as judged by humans (blue) and the regressor (red). These two lines are compared with gold texts, again assessed by humans (yellow) and the regressor (green).

Comparing the models and the humans as assessed by humans (lines blue and yellow) we see that while for short sentences humans perceive the generated and the natural texts equally human-like, this changes substantially for longer fragments. At length 40, we observe the largest gap in perception between the models and the natural texts, with the latter being perceived much more human-like.

In terms of machine-based evaluation (lines red and green), the behaviour of the BERT regressor on human data is very similar to the human judgements (line green vs yellow). Although the two curves are similar also for the texts generated by the models, the regressor here overestimates as human-produced texts that are actually machine generated (line red vs blue). This is potentially due

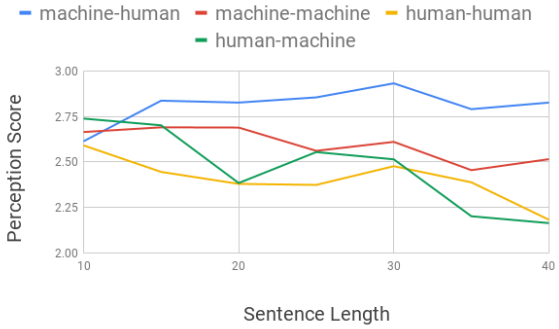

Figure 1: Average perception scores for human vs machine generated texts as assessed by humans and our regressor. In legend: <producer-assessor>. Machine scores are averaged across the three models.

to the fact that GePpeTto-F and GePpeTto-R use the same (human labelled) training data for fine-tuning which is used to train the regressor model. This phenomenon appears exacerbated with longer texts, as the blue and red lines are more distant after length 20.[7] This behaviour of the regressor is also reflected by its scores being more compressed towards the middle. Indeed, the average standard deviations in Table 3, show higher variability in human judgements than in the regressor's assessment. In Table 4 same examples of generated sentences together with their scores are reported.

5 Linguistic Analysis

We ran a linguistic analysis over the human and the generated text using Profiling-UD (Brunato et al., 2020), a tool that extracts linguistic features of varying complexity, ranging from raw text aspects, such as average length of words and sentences, to lexical, morpho-syntactic, and syntactic properties. In particular, we study (i) which features characterise the most humanly-perceived texts in the training data, independently of who generated them; (ii) the difference between human-produced texts and those generated by our best model (GePpeTto-F) in the test set when they are perceived as human.[8]

Regarding (i), the features that most correlate with a text being perceived as human have to do with sentence length and complexity. For example, the longer the sentence or the clauses therein, or the longer and deeper the syntactic links, the more humanly-perceived is the text. On the other side of the spectrum, linguistic features associated to texts

[7] The detailed tables in the Appendix further show this divergence with specific scores per model.

[8] Findings summarised; detailed correlations in Appendix.

model	output	human-score	regressor-score
Human	La ex Chiesa di Santa Caterina del Monte di Pietà era una chiesa cattolica che si trova ad Alcamo, in provincia di Trapani. (The former Church of Santa Caterina del Monte di Pietà was a Catholic church located in Alcamo, in the province of Trapani.)	1.71	1.88
GePpeTto-F	La nuova sede fu inaugurata il 19 luglio 1885 e inaugurata ufficialmente il 30 novembre 1889, giorno in cui fu completata la facciata. (The new headquarters were inaugurated on July 19, 1885 and officially inaugurated on November 30, 1889, the day the facade was completed.)	1.86	2.34
GePpeTto-R	La casa si trova in una posizione favorevole all'espansione del mercato e, in alcuni casi, alla costruzione di tende per bambini. (The house is in a favorable position for the expansion of the market and, in some cases, for the construction of children's tents.)	3.14	2.68
GePpeTto	La squadra era composta di due squadre, una delle quali era la "Rhodesliga" con il termine del "Propaganda Fiumana". (The team was made up of two teams, one of which was the "Rhodesliga" with the term of "Propaganda Fiumana".)	3.15	3.07

Table 4: Sample model outputs and their sentence-level score. Prompt: "La" ("The$_{[feminine]}$").

judged as machine-generated are heavy presence of punctuation and of interjections and symbols.

For (ii), we zoom in on humanly-perceived texts only, but looking at the source that generated them. For human texts, length and complexity are still the relevant features for being perceived as human; these are proxied by complex verbal structures characterised by auxiliaries, use of past tense, number of main predicates in a sentence. For the generated texts, instead, we observe that both those characteristics that are similar to the human texts, such as the use of the indicative mood and finite tenses, as well as those more specific to machine-generated texts, such as a low density of subordinate clauses and shorter sentences, are simpler structures where it is more likely that the machine does not incur evident mistakes: it is easier for the model to produce human looking sentences if they are kept short and simple. With longer sentences the model struggles to ensure semantic and pragmatic coherence, two aspects that most likely require further and more complex modelling beyond simple fine-tuning.

6 Conclusions

We elicited judgements on the human-likeness of gold and generated Italian texts and used these judgements to fine-tune a pre-trained GPT-2 model to push it to produce more human-like texts. Our evaluation shows that people indeed find the output of the fine-tuned model more human-like than that of the basic one. Contextually, we show that our proposed automatic evaluation correlates well with human judgements, and it is therefore a reliable strategy that can be applied in absence of subjects.

An analysis of linguistic features reveals that while complexity is associated with human-likeness in gold data, simplicity is a key feature of artificial texts that are assessed as human-like, perhaps because simpler texts are less prone to expose machine behaviour.

Future work will include an expansion of the perception-labelled data to (i) assess training size in fine-tuning, and (ii) perform a finer-grained analysis correlating assessments to different text genres and subject demographics.

Impact Statement

All work that automatically generates text could unfortunately be used maliciously. While we cannot fully prevent such uses once our models are made public, we do hope that writing about risks explicitly and also raising awareness of this possibility in the general public are ways to contain the effects of potential harmful uses. We are open to any discussion and suggestions to minimise such risks. The contributors of human judgements elicited for this work have been fairly compensated.

Acknowledgements

We would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high performance computing cluster. We are also grateful to the anonymous GEM reviewers whose comments contributed to improving this paper.

References

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2020. Profiling-ud: a tool for linguistic profiling of texts. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 7145–7151.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2020. Geppetto carves italian into a language model. In *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021*, volume 2769 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Hongyu Gong, Suma Bhat, Lingfei Wu, JinJun Xiong, and Wen-mei Hwu. 2019. Reinforcement learning based text style transfer without parallel training corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3168–3180.

Tatsunori B. Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. *CoRR*, abs/1904.02792.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822, Online. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.

Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, and Y-Lan Boureau. 2019. Multiple-attribute text rewriting. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Fuli Luo, Peng Li, Jie Zhou, Pengcheng Yang, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. A dual reinforcement learning framework for unsupervised text style transfer. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 5116–5122.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Abhilasha Sancheti, Kundan Krishna, Balaji Vasan Srinivasan, and Anandhavelu Natarajan. 2020. Reinforced rewards framework for text style transfer. In *Advances in Information Retrieval*, pages 545–560.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Appendix

This Appendix contains:

- detailed results of human and machine evaluation for gold and all models' data (Tables A.1–A.2), expanding the compressed results shown in Table 2 in the main paper.
- details of linguistic features (correlated with human and machine perception, Tables A3–A4) which are discussed in Section 5 in the main paper.

	Length							
Tipo	10	15	20	25	30	35	40	AVG
GePpeTto	2.80	2.83	3.05	2.89	3.08	2.55	2.77	2.85 (0.83)
GePpeTto-F	2.44	2.68	2.57	2.85	2.74	2.97	2.93	2.74 (0.83)
GePpeTto-R	2.61	3.01	2.87	2.83	2.97	2.85	2.78	2.84 (0.87)
Human	2.59	2.45	2.38	2.37	2.48	2.39	2.18	2.41 (0.77)
avg	2.61	2.74	2.72	2.74	2.82	2.69	2.67	2.71 (0.85)

Table A.1: Average scores for each system grouped by sentence length as assigned by humans on the test set.

	Length							
Tipo	10	15	20	25	30	35	40	AVG
GePpeTto	2.79	2.78	2.88	2.80	2.76	2.53	2.68	2.74 (0.71)
GePpeTto-F	2.53	2.62	2.52	2.44	2.44	2.46	2.43	2.49 (0.55)
GePpeTto-R	2.68	2.67	2.67	2.45	2.63	2.38	2.44	2.56 (0.57)
Human	2.74	2.70	2.38	2.55	2.51	2.20	2.16	2.47 (0.66)
avg	2.68	2.69	2.61	2.56	2.59	2.39	2.43	2.57 (0.63)

Table A.2: Average scores for each system grouped by sentence length as assigned by the BERT based regressor on the test set.

Human Texts		Generated Texts	
Feature	Correlation (p-values)	Feature	Correlation (p-values)
n_tokens	-0.2 (7.34e-14)	upos_dist_NOUN	-0.15 (1.08e-08)
avg_max_links_len	-0.19 (1.85e-13)	dep_dist_compound	-0.13 (2.17e-06)
max_links_len	-0.18 (2.52e-11)	subj_pre	-0.1 (8.90e-05)
avg_max_depth	-0.17 (1.86e-10)	prep_dist_1	-0.09 (6.15e-04)
avg_links_len	-0.13 (2.43e-06)	avg_prepositional_chain_len	-0.09 (7.12e-04)
avg_token_per_clause	-0.12 (1.20e-05)	n_prepositional_chains	-0.09 (7.53e-04)
upos_dist_X	-0.1 (1.20e-04)	n_tokens	-0.09 (8.45e-04)
dep_dist_goeswith	-0.1 (1.43e-04)	dep_dist_amod	-0.08 (2.37e-03)
verbal_head_per_sent	-0.1 (3.50e-04)	upos_dist_ADJ	-0.08 (2.39e-03)
subj_pre	-0.09 (4.07e-04)	dep_dist_nsubj	-0.08 (4.03e-03)
verbal_root_perc	-0.09 (6.48e-04)	avg_max_depth	-0.08 (4.06e-03)
avg_verb_edges	-0.09 (1.03e-03)	avg_token_per_clause	-0.08 (4.19e-03)
obj_post	-0.09 (1.04e-03)	dep_dist_case	-0.07 (8.76e-03)
verbs_num_pers_dist_+	-0.08 (1.54e-03)	max_links_len	-0.07 (9.10e-03)
dep_dist_det	-0.08 (1.58e-03)	verbs_form_dist_Inf	-0.07 (9.78e-03)
		...	
dep_dist_iobj	0.04 (9.89e-02)	dep_dist_nmod:tmod	0.04 (1.15e-01)
dep_dist_appos	0.05 (7.38e-02)	verb_edges_dist_1	0.04 (1.10e-01)
dep_dist_advcl	0.05 (7.29e-02)	dep_dist_advmod	0.04 (1.01e-01)
dep_dist_flat	0.06 (1.87e-02)	aux_mood_dist_Imp	0.05 (8.17e-02)
lexical_density	0.06 (1.84e-02)	upos_dist_CCONJ	0.06 (2.98e-02)
subordinate_dist_3	0.06 (1.56e-02)	upos_dist_PROPN	0.06 (2.35e-02)
dep_dist_nmod:tmod	0.07 (1.38e-02)	dep_dist_discourse	0.08 (4.42e-03)
aux_form_dist_Inf	0.07 (9.01e-03)	dep_dist_appos	0.08 (3.43e-03)
dep_dist_nummod	0.08 (2.63e-03)	upos_dist_INTJ	0.08 (2.43e-03)
upos_dist_PROPN	0.11 (4.98e-05)	dep_dist_conj	0.08 (1.61e-03)
upos_dist_NUM	0.12 (3.36e-06)	verbs_form_dist_Ger	0.09 (1.01e-03)
upos_dist_PUNCT	0.13 (2.08e-06)	upos_dist_SYM	0.11 (4.11e-05)
upos_dist_SYM	0.13 (1.26e-06)	dep_dist_root	0.11 (1.70e-05)
dep_dist_punct	0.14 (1.65e-07)	upos_dist_PUNCT	0.25 (1.31e-21)
dep_dist_root	0.26 (2.07e-22)	dep_dist_punct	0.25 (4.71e-22)

Table A.3: Linguistic features in training data. Generated = GePpeTto base

Human Texts		Generated Texts	
Feature	Correlation (p-values)	Feature	Correlation (p-values)
verbal_root_perc	-0.28 (9.25e-08)	principal_proposition_dist	-0.2 (1.33e-04)
verbs_tense_dist_Past	-0.21 (6.34e-05)	dep_dist_nsubj:pass	-0.19 (3.55e-04)
upos_dist_DET	-0.18 (8.41e-04)	dep_dist_aux:pass	-0.18 (5.71e-04)
dep_dist_det	-0.17 (1.19e-03)	dep_dist_root	-0.18 (7.22e-04)
aux_form_dist_Fin	-0.17 (1.33e-03)	aux_mood_dist_Ind	-0.18 (7.45e-04)
upos_dist_AUX	-0.17 (1.51e-03)	aux_form_dist_Fin	-0.17 (1.94e-03)
aux_num_pers_dist_Sing+3	-0.17 (1.65e-03)	aux_tense_dist_Past	-0.16 (1.97e-03)
verbal_head_per_sent	-0.17 (1.79e-03)	aux_num_pers_dist_Sing+3	-0.16 (2.87e-03)
aux_mood_dist_Ind	-0.16 (2.11e-03)	dep_dist_obl:agent	-0.16 (3.25e-03)
dep_dist_obl	-0.16 (2.31e-03)	verbal_root_perc	-0.14 (7.63e-03)
dep_dist_expl	-0.16 (2.45e-03)	dep_dist_flat	-0.13 (1.20e-02)
dep_dist_case	-0.14 (6.99e-03)	dep_dist_det	-0.13 (1.57e-02)
aux_tense_dist_Past	-0.14 (7.98e-03)	lexical_density	-0.12 (2.06e-02)
dep_dist_cop	-0.13 (1.22e-02)	upos_dist_AUX	-0.12 (2.74e-02)
upos_dist_ADP	-0.13 (1.55e-02)	verb_edges_dist_5	-0.11 (4.33e-02)
		...	
dep_dist_flat:name	0.1 (5.16e-02)	n_prepositional_chains	0.12 (2.02e-02)
verbs_tense_dist_Pres	0.11 (3.77e-02)	verbs_num_pers_dist_Plur+3	0.13 (1.59e-02)
verbs_form_dist_Inf	0.11 (3.54e-02)	dep_dist_punct	0.13 (1.48e-02)
char_per_tok	0.12 (2.98e-02)	upos_dist_PUNCT	0.13 (1.40e-02)
dep_dist_compound	0.12 (2.40e-02)	dep_dist_nummod	0.15 (4.12e-03)
dep_dist_root	0.14 (1.08e-02)	dep_dist_conj	0.15 (3.76e-03)
upos_dist_PUNCT	0.14 (9.63e-03)	upos_dist_PRON	0.16 (3.56e-03)
dep_dist_punct	0.14 (9.63e-03)	upos_dist_SYM	0.16 (2.19e-03)
upos_dist_PROPN	0.15 (6.28e-03)	dep_dist_acl:relcl	0.17 (1.36e-03)
dep_dist_nmod	0.17 (1.81e-03)	dep_dist_appos	0.17 (1.29e-03)
upos_dist_SYM	0.17 (1.63e-03)	n_tokens	0.19 (4.73e-04)
dep_dist_nummod	0.17 (1.17e-03)	tokens_per_sent	0.19 (4.73e-04)
lexical_density	0.17 (1.12e-03)	avg_links_len	0.25 (1.94e-06)
dep_dist_flat	0.22 (2.46e-05)	avg_max_links_len	0.26 (1.02e-06)
upos_dist_NUM	0.25 (2.08e-06)	max_links_len	0.26 (1.02e-06)

Table A.4: Linguistic features on test data. Generated = GePpeTto-F.