INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Pragmatic Pertinence: A Learnable Confidence Metric to Assess the Subjective Quality of LM-Generated Text

Jerome R. Bellegarda

Etsy Inc., Brooklyn, New York 11201, USA*

jerome@etsy.com

Abstract

To be perceived as trustworthy, artificially generated text must be sufficiently congruent with the available discourse history. Pre-trained language models (LMs) operating in generative mode are capable of predicting locally coherent phrases, but those do not always reflect salient syntactic, semantic, or pragmatic facets of prior content. This paper introduces a learnable evaluation metric to assess the pragmatic *pertinence* of LM-generated text for a given history. Pertinence is closely aligned with qualitative human judgments of acceptability, thereby emerging as a blend of sensibleness and specificity. Experiments conducted across different domains using different learning architectures show that this approach circumvents the issue of multiple valid ground-truths, while providing a reliable quantitative ranking of generated text completion candidates in context. Pertinence scoring could thus prove useful for the detection of hallucinations.

Index Terms: natural language generation, text completion, discourse congruence, evaluation metrics, human judgments, hallucination detection

1. Introduction

In many applications, ranging from speech/handwriting recognition (e.g., [1]) to keyboard input prediction (e.g., [2]), transformer-based neural language models (LMs) have become the *de facto* standard for estimating the relative likelihood of predicted word sequences [3]. These LMs are inherently generative; once trained, they can be used to generate phrases of arbitrary length by feeding their previous outputs back into the model [4], [5]. Yet, it is generally acknowledged that state-of-the-art implementations comprising a large number of parameters (e.g., 175 billion in the case of GPT-3 [6], 540 billion in the case of PaLM [7]) have difficulties generating consistently high-quality natural language text across a large enough variety of domains and contexts [8], [9].

Among other drawbacks, predicted texts are occasionally bland, incoherent, repetitive, nonsensical, and/or unfaithful to the provided source input [10]. The risk of producing "hallucinations" has been a limiting factor in a number of applications [11]. And compared to human experts, even recent models fine-tuned from human feedback like ChatGPT [12] sometimes perform poorly in terms of helpfulness in question-answering [13].

Quantifying such shortcomings in natural language generation requires a deep understanding of discourse content, which automatic assessment measures typically lack. Objective measures such as perplexity or key saving ratio (used in, e.g., predictive typing) are too word-centric to be satisfactory at the phrase level across diverse domains. Similar metrics imported from areas such as machine translation or summarization [14] (such as BLEU, ROUGE, and CIDEr) calculate distances between generated text and original ground-truth, but sidestep the issue of multiple valid ground-truths.

As for subjective measures like the multi-turn Likert score (used to evaluate, e.g., conversational systems), they are designed to gauge complex attributes (such as naturalness, originality, appropriateness, engagingness, etc.) by having humans evaluate things like multi-turn consistency, repetition avoidance, etc. But *a posteriori* human judgments are slow, tedious, costly, and often yield comparisons that are not statistically significant [15]. In addition, they are clearly unsuitable when it comes to inline assessment.

This observation has sparked interest in data-driven confidence metrics capable of acting as inline proxies for such subjective measures. This paper proposes to learn such a metric directly in embedding space through a classifier trained to emulate qualitative human judgments of pertinence. For a given linguistic discourse history, the classifier determines the extent to which the associated generated text constitutes an acceptable completion. The outcome is a data-driven assessment of the pragmatic quality of generated text.

To summarize the contributions of this work:

- we introduce the concept of a learnable discriminative measure to assess qualitative congruence with linguistic discourse history;
- we establish a new task/benchmark: predicting the quality of LM-generated text based on mimicking elicited human judgments of pertinence;
- we describe how to design/collect/annotate a suitable training corpus of qualitative human judgments to train the pertinence classifier;
- experiments show that pertinence prediction is robust across domains and across classifier architectures.

2. Related work

There have been numerous efforts to construct automatic assessment measures from sentence embeddings (based on, e.g., BERT [16]) for both context and prediction. For example, for each turn in a dialog, it is possible to compute the distance in embedding space between the sentence embeddings associated with human prompt and

* This work was performed while the author was still with Apple, Cupertino, California, with help from Bishal Barman.

10.21437/Interspeech.2023-180

chatbot answer. Unfortunately, such metrics do not correlate well with human evaluations [17], most likely due to the heavy-tailed distribution of natural language and the varying strength of discourse-imposed constraints. Besides, common embedding distance measures (such as the cosine distance) are not inherently discriminative, so separating "good" from "bad" predicted material often comes down to somewhat arbitrary, hard-to-justify empirical thresholds in embedding space.

The need to properly calibrate such thresholds has underscored the challenge of gathering the necessary human annotations. Such annotations tend to focus on characteristics of natural language generation that reflect human conversational abilities. For example, in chatbot evaluation, annotators answer questions such as: "*which speaker is more engaging?*" or "*which speaker is more knowledgeable?*" [15]. The average of *sensibleness* and *specificity* (SSA) is often used as a proxy for "human-like"-ness that specifically penalizes chatbots for producing generic responses [18]. Further human-like traits may even be brought to bear, like personality, humor, empathy, and related interpersonal qualities. But scoring such fine-grained attributes tends to place a heavy burden on annotators, with deleterious consequences in terms of inter-annotator agreement.

In contrast, this paper targets the pragmatic quality of generated text conditioned solely on linguistic discourse history. By nature this assessement involves a blend of sensibleness and specificity, intrinsically anchored to qualitative human judgments of acceptability.

3. Corpus design and construction

The first step is therefore to elicit appropriate human judgments of pertinence, preferrably in a domain-agnostic way. But evaluating the quality of any textual generation requires care, because the complexity of natural language normally translates into more than one way to achieve congruence with the discourse context.

To illustrate the point, Fig. 1 below shows an example discourse history Z and 6 possible completions denoted by W_1 to W_6. It is easy to eliminate W_1, W_3, and W_5 as not pertinent, albeit for different reasons: W_1 due to domain mismatch, W_3 because "a *waste of time*" contradicts "*marvelous experience*", and W_5 because it is not a well-formed sentence. But it is pointless to select which of W_2, W_4, and W_6 is most pertinent. In the absence of any further information, all three completions can be viewed as alternative ground truths, all perfectly valid

Z	Hiking on the John Muir trail last July was such a marvelous experience!	
W_1	*It truly was* the best chicken potpie I ever had.	0
W_2	*It truly was* the highlight of the summer.	1
W_3	*It truly was* a waste of time.	0
W_4	*It truly was* a once in a lifetime trip!	1
W_5	*It truly was* a great way to bond with	0
W_6	*It truly was* a great way to bond with my son.	1

Figure 1: *Example of discourse history (Z) and possible predicted completions (W_1 to W_6). Scores in the last column reflect ideal pertinence.*

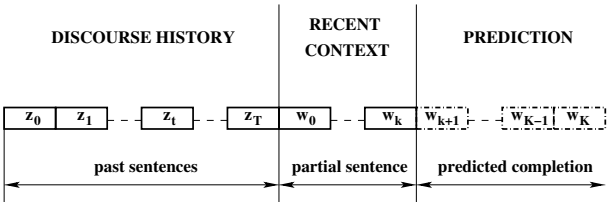

Figure 2: *Past, partial, and predicted word sequences.*

given the available context. Thus, quality assessment of textual generation inherently depends on multiple potential ground truths.

This observation provides an incentive to assess the subjective quality of LM-generated text based on mimicking qualitative human judgments of completion acceptability as a whole. Arguably, humans arrive at such judgments by applying a subtle blend of sensibleness and specificity scoring, with no attempt to artificially disentangle the two dimensions. In other words, they seek to assess the *pragmatic pertinence* of every potential completion given the available context. In Fig. 1, for example, ideal pertinence scores would be 0 for W_1, W_3, and W_5 and 1 for W_2, W_4, and W_6 (as shown in the last column).

Accordingly, the notion of pertinence emerges as a reasonable measure for ranking text completion candidates at inference time. For metric learning purposes, we need to gather a set of typical human judgments observed in various situations. To ensure diversity, human judgments are solicited for a wide array of discourses across multiple domains, ranging from short stories to social media conversations. The basic elements of the annotation task are depicted in Fig. 2. By design, we keep the task as simple as possible: to assess whether or not the predicted completion is acceptable given the available discourse history.

We begin by extracting short paragraphs from various text sources, from where to draw suitable (discourse, completion) pairs. The discourse history $Z = z_0 \dots z_T$ consists of 4 to 8 short sentences that define the general fabric of discourse. The partial sentence $w_0 \dots w_k$ consists of the beginning of the next sentence $W = w_0 \dots w_K$, which we truncate after a variable number of words $1 < k < 5$ to form the recent context. The purpose of the recent context is to focus the predictions generated by the different LMs, so as to make the task of the human annotators more manageable.

N annotators are then asked to rate $K = 5$ different completions for that recent context. Only one of these completions (the "golden completion") results in the original sentence W observed in the extracted paragraph/conversation. The other $K - 1$ completions are generated by different LMs of varying quality, to provide a range of plausibility and specificity to the discourse. For every (discourse, completion) pair, annotators thus provide ideal pertinence values (1 for pertinent and 0 for non pertinent, as in Fig. 1). Instead of hard scores, we also experimented with soft scores based on the annotator's perceived degree of pertinence, but found that inter-annotator agreement suffered as a result. With hard scores, we obtain Cohen's $\kappa = 0.42$, which is fairly typical of such tasks [19]. Note that the same pool of annotators is used across all domains considered.

We perform quality control by examining the distri-

Z	Allie wanted to paint her nails. She chose a bright red color. But the nail polish was very thick. It needed a long time to dry.	Cosine similarity	Pragmatic pertinence
W_1	*Allie had to* wait for half an hour.	0.5	0.56
W_2	*Allie had to* paint her nails again with a deeper red.	0.6	0.37
W_3	*Allie had to* take the bus to work the next day.	0.2	0.44
W_4	*Allie had to* ride a motorcycle with a helmet.	0.3	0.42

Figure 3: *Predicted completions W_1 to W_4 for discourse history Z, ordered in decreasing order of human-judged pertinence. The last two columns reflect automated similarity scores between discourse and completions.*

bution of ratings thus produced for every discourse history. First we define the human-judged pertinence (HJP) of every completion as the proportion of annotators who are in agreement that the completion is pertinent to the overall context. Then we compute the average HJP value across all completions for a given history. An average HJP value around 1 suggests that only the golden completion is deemed pertinent for this history, while an average completion close to K indicates that most completions are deemed acceptable. We aim to ideally include only those histories which are collectively associated with an average HJP around $K/2$, in order to ensure a proper balance of difficulty given the set of LMs considered.

4. Learnable confidence measure

Given two textual inputs (such as discourse history and predicted completion), automated testing for textual congruence typically involves some measure of statistical dependence between pairs of sentence embedding vectors and human judgments for the same pairs. A measure commonly used is the Pearson correlation coefficient, but it only assesses linear relationships, and is known to be sensitive to outliers [20]. Calibration success is therefore not guaranteed. In addition, it is heavily dependent on the given embedding space and the choice of similarity measure adopted.

To circumvent such shortcomings, we adopt a learned similarity strategy, where both sentence embeddings and similarity measure are learned from data, and the evaluation system directly predicts pertinence as a proxy for human judgment. We empirically choose the architecture depicted in Fig. 4 below, where transformer encoder stacks are used to extract embedding vectors (p and q, respectively) for both discourse history and predicted text, and the similarity between them is calculated via a transformer decoder stack. This model can be viewed as a

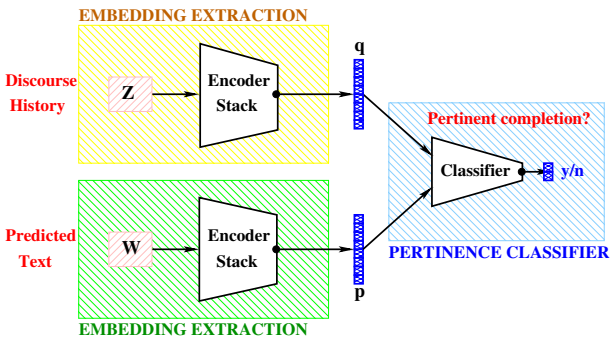

Figure 4: *Neural pertinence predictor architecture.*

binary classifier trained on ideal pertinence values (1 for pertinent and 0 for non pertinent, cf. Fig. 1) associated with (discourse, completion) pairs. The outcome is a numerical pertinence score which reflects how good a particular completion is given a particular history.

Fig. 3 above illustrates the outcome for a given discourse history and 4 different predicted completions, ordered in decreasing order of human-judged pertinence. Two automated similarity scores between discourse and completions are shown: the conventional cosine similarity score serving as baseline, and the pertinence score learned as described above. Neither score is perfectly aligned with the human ranking, but the top-rated pertinence is in agreement with the preferred human completion, while the top cosine score is not. Also note that, while W_2 has a large lexical overlap with the context (which likely explains the high cosine score), pragmatically it is a rather weak completion, as correctly reflected in the lower pertinence score.

Fig. 3 underscores the ineffectuality of using non-discriminative similarity measures (like cosine) as baselines for the problem considered here (i.e., mimicking human judgments). They do not constitute convincing baselines precisely because they do not correlate well with human judgments, as discussed at some length in [17].

5. Experimental setup

For training the neural pertinence predictor of Fig. 4, we gather real data from four sources with very different characteristics in terms of both style and homogeneity: conversational data collected from diverse scenarios, commonsense stories originally put together for an evaluation of narrative structure learning (Story Cloze Test [21], [22]), Reddit threads on about 10 topics ranging from Art to Travel, and high-frequency n-grams collected via differential privacy.

For the first three sources, every conversation, story, and thread is split into constituent parts as per Fig. 2, with the golden completion as one possible completion. We then generate alternative completions from the context using $K-1=4$ different LMs of varying strength, such that by design we get a mix of "good" and "bad" completions. The outcome is a tuple formed by the context and the resulting set of completion candidates. And finally we submit each (randomized) tuple to $N=5$ different crowd workers so they can render their judgments of pertinence. This process resulted in about 2500 labelled tuples for the conversational data, 55000 for the stories, and 50000 for the Reddit threads.

Since, technically, high-frequency n-grams are all pertinent, we leverage the last source to encourage the learned metric to favor proper syntax. We separate syn-

Table 1: *Accuracy results across domains.*

Training/Test	Stories	Reddit
Stories	79.2 %	73.6 %
Reddit	73.0 %	81.3 %
All	79.5 %	81.6 %

tactically complete predictions from others by partitioning the data into two bins: terminating n-grams, which occur at the end of a sentence, and non-terminating n-grams, which do not. We then label any terminating n-gram as pertinent and any non-terminating n-gram as non-pertinent. We use $n = 8$ throughout, with the first and last 4 words taken as context and completion, respectively. This process led to about 4500 labelled tuples.

The collection of all annotated tuples is then split into 80% training, 10% validation, and 10% test sets. To experiment with a diversity of classifiers, we fine-tune different pre-trained LMs, and use top-1 classification accuracy as evaluation metric. In all cases we use standard defaults for the various training protocols to facilitate reproducibility. Given its poor performance, we only ran the baseline cosine similarity metric on Stories: as expected, its top-1 classification accuracy was less than 40%, making it essentially unusable in practice.

6. Results and discussion

Table 1 reports results obtained across two domains: Stories and Reddit. Table 2 reports results obtained across eight pre-trained architectures.

Table 1 shows that pertinence prediction is fairly robust across domains, and that accuracy increases when training on multiple domains. The fact that only a small drop in accuracy is observed when pertinence trained in one domain is applied to another domain suggests that it is possible to mimic human judgments of pertinence in a largely domain-agnostic way. Table 2 shows that it matters little which pre-trained architecture is used for fine-tuning. In addition to indicating that the choice of a specific model architecture is not material, Table 2 also suggests that fine-tuning only a relatively small number of parameters is sufficient to successfully rank text completion candidates in context.

To illustrate real-word usage, Figs. 5 and 6 show two actual predictions scored using the confidence measure. In each case, the discourse history Z is drawn from the Stories sub-corpus ([21], [22]), and the prediction W_1 corresponds to the top completion generated by a ∼30MB transformer-based LM running on-device. Fig. 5 illustrates a case of high pragmatic pertinence: W_1 is generally congruent with the context, which is reflected in the high confidence score of 0.89. In contrast, Fig. 6 illustrates a case of low pragmatic pertinence: here W_1 does not make sense given the available context, which is reflected in the low confidence score of 0.28.

Table 2: *Accuracy results across architectures.*

Architecture/Test	Stories	Reddit	N-grams
BERT	78.8 %	79.2 %	97.7 %
RoBERTa	81.0 %	80.5 %	94.3 %
DistillBERT	78.9 %	79.3 %	96.1 %
DistillRoBERTa	79.3 %	82.8 %	96.1 %
SqueezeBERT	78.8 %	80.2 %	95.6 %
MobileBERT	79.2 %	82.2 %	96.3 %
ConvBERT	78.3 %	79.6 %	94.3 %
ALBERT	79.7 %	80.8 %	93.6 %

These experiments suggest that the proposed framework is viable to encapsulate human judgments of pertinence. They also bode well for the ability of pertinence scoring to automatically assess the subjective quality of LM-generated text. By enforcing congruence with the discourse history, this confidence measure may ultimately help surface more natural, concise, and/or expressive text completion candidates. It could also play a crucial role in the detection and reduction of hallucinations, for example by providing an alternative to costly human feedback in reinforcement learning approaches (cf. [23]).

7. Conclusion

Pragmatic pertinence is a learnable metric blending sensibleness and specificity into a single automated measure of quality. It is trained to mimic qualitative human judgments of acceptability for a given LM-generated text completion in a given discourse history. It thus serves as a confidence measure to re-rank a list of text completion candidates in a given context. Training data is composed of human judgments elicited for a variety of (discourse, completion) pairs in multiple domains. The metric operates in a domain-agnostic way and can be fine-tuned from a variety of pre-trained LMs using a comparatively small number of human judgments.

We observed across different domains and different learning architectures that this approach helps quantify the extent to which natural language generation reflects syntactic, semantic, and pragmatic facets of discourse, while circumventing the issue of multiple valid ground-truths. The confidence measure can therefore be advantageously adopted in a wide variety of LM-driven text generation applications to promote higher-quality and more trustworthy outcomes.

Z	Valerie hired John to install her kitchen cabinets. John was always late getting to work. John did not finish installing the cabinets on time. Valerie was not happy with John's work.	
W_1	*She told him* that he was going to have to come back to the house.	0.89

Figure 5: *Example of high pertinence score prediction.*

Z	Sarah walked into library. The metal detector made a sound as soon as she walked through it. A security guard pulled her to the side of the detector. He ran a quick search.	
W_1	*He was able* to find the metal detector on the floor of the library.	0.28

Figure 6: *Example of low pertinence score prediction.*

8. References

[1] Y. Chherawala, H. Dolfing, R. Dixon, and J.R. Bellegarda, "Large-Scale Handwritten Chinese Character Recognition for Mobile Devices," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc.*, Barcelona, Spain, May 2020.

[2] A. Mehra, J.R. Bellegarda, O. Bapat, P. Lal, and X. Wang, "Leveraging GANs to Improve Continuous Path Keyboard Input Models," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Proc.*, Barcelona, Spain, May 2020.

[3] 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 Neural Information Processing Systems*, Vol. 30, pp. 5998–6008, 2017.

[4] A. Radford, K. Narasimhan, T. Salimans and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," OpenAI Assets Research Covers, 2018.

[5] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q.V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context," *arXiv:1901.02860v3*, June 2019.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language Models are Few-Shot Learners," *arXiv:2005.14165v4*, July 2020.

[7] A. Chowdhery *et al.*, "PaLM: Scaling Language Modeling with Pathways," *arXiv:2204.02311*, 2022.

[8] A. Rohrbach, L.A. Hendricks, K. Burns, T. Darrell, and K. Saenko, "Object Hallucination in Image Captioning," in *Proc. EMNLP*, 2018.

[9] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The Curious Case of Neural Text Degeneration," in *Int. Conf. Learning Representations*, 2019.

[10] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung, "Survey of Hallucination in Natural Language Generation," *arXiv:2202.03629*, Feb. 2022.

[11] T.A. Chang and B.K. Bergen, "Word Acquisition in Neural Language Models," *Trans. Assoc. Computational Linguistics*, 10: 1–16, Jan. 2022.

[12] OpenAI, "ChatGPT: Optimizing Language Models for Dialogue," https://openai.com/blog/chatgpt, Nov. 2022.

[13] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, "How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection," *arXiv:2301.07597v1*, Jan. 2023.

[14] C. Hori, J. Perez, R. Higasinaka, T. Hori, Y.L. Boureau, "Overview of the Sixth Dialog System Technology Challenge: DSTC6," *Computer Speech & Language*, Vol. 55, 2018.

[15] M. Li, J. Weston, and S. Roller, "Acute-Eval: Improved Dialogue Evaluation with Optimized Questions and Multi-Turn Comparisons," *arXiv:1909.03087v1*, Sept. 2019.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. Conf. North American Chapter Assoc. Comput. Linguistics*, Minneapolis, MN, pp. 4171-4186, 2019.

[17] C.-W. Liu, R. Lowe, I.V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation", *Computing Research Repository (CoRR)*, abs/1603.08023, 2016.

[18] D. Adiwardana, M.-T. Luong, D.R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q.V. Le, "Towards a Human-like Open-Domain Chatbot," *arXiv:2001.09977v2*, Jan. 2020.

[19] M.L. McHugh, "Interrater Reliability: The Kappa Statistic," *Biochemia Medica*, Vol. 22, No. 3, pp. 276-282, 2012.

[20] S. Eger, A. Rücklé, and I. Gurevych, "Pitfalls in the Evaluation of Sentence Embeddings," in *Proc. 4th Workshop on Repres. Learning for NLP (RepL4NLP-2019)*, pp. 55–60, Florence, Italy, Aug. 2019.

[21] N. Mostafazadeh, M. Roth, A. Louis, N. Chambers, and J. Allen, "LSDSem 2017 Shared Task: The Story Cloze Test," in *Proc. Workshop Linking Models Lexical, Sentential and Discourse-level Semantics*, pp. 46–51, 2017.

[22] S. Srinivasan, R. Arora, and M. Riedl, "A Simple and Effective Approach to the Story Cloze Test," in *Proc. Conf. North American Chapter Assoc. Computational Linguistics: Human Language Technologies*, Vol. 2, pp. 92–96, 2018.

[23] B. Zhu, J. Jiao, and M.I. Jordan, "Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons," *arXiv:2301.11270v3*, March 2023.