Information Value: Measuring Utterance Predictability as Distance from Plausible Alternatives*

Mario Giulianelli[◁†] **Sarenne Wallbridge**[○†] **Raquel Fernández**[◇]
[◁]ETH Zürich, Department of Computer Science
[○]University of Edinburgh, Centre for Speech Technology Research
[◇]University of Amsterdam, Institute for Logic, Language and Computation
mgiulianelli@inf.ethz.ch s.wallbridge@ed.ac.uk raquel.fernandez@uva.nl

Introduction

Measuring the amount of information carried by a linguistic signal is fundamental to the computational modelling of language processing. Such quantifications are used in psycholinguistic and neurobiological models of human language processing (Futrell and Levy, 2017; Armeni et al., 2017), algorithmic linguistic theories of utterance acceptability (Lau et al., 2017), to study the processing mechanisms of neural language models (e.g., Futrell et al., 2019; Sinclair et al., 2022), to power sampling algorithms for natural language generation (Wei et al., 2021; Meister et al., 2023), and as a learning and evaluation criterion. The amount of information carried by a linguistic signal is intrinsically related to its predictability, as summarised in the definition of the *surprisal* of a unit u (Shannon, 1948), perhaps the most widely used measure of information: $I(u) = -\log_2 p(u)$. Predictable units carry low amounts of information—i.e., low surprisal—as they are already expected to occur given the context in which they are produced. Conversely, unexpected units carry higher surprisal.

Estimation of the surprisal of an utterance in the space of natural language strings would require computing probabilities over a high-dimensional, structured, and ultimately unbounded event space. It is thus common to resort to chaining token-level surprisal estimates, nowadays typically obtained from neural language models (Giulianelli and Fernández, 2021; Meister et al., 2021; Wallbridge et al., 2022). However, such token-level autoregressive approximations of utterance probability conflate different dimensions of predictability (see, e.g., Arehalli et al., 2022; Kuhn et al., 2023), which makes it difficult to appreciate whether the information carried by an utterance is a result, for example, of the unexpectedness of its lexical material, syntactic arrangements, semantic content, or speech act type.

Alternative-Based Information Value

We propose an intuitive characterisation of the information carried by utterances which computes predictability over the space of full utterances and explicitly models multiple dimensions of uncertainty, thereby offering greater interpretability of utterance predictability estimates. Given a linguistic context, x, the *information value* of an utterance y is defined as the distribution of distances between y and the set of contextually expected alternatives A_x, measured with a distance metric d:

$$I(Y=y|X=x) := d(y, A_x) \tag{1}$$

Distributions skewed towards large distances indicate that y differs substantially from expected utterances, and thus that y is a surprising contribution to discourse, with high information value.

In practice, computing the information value of an utterance requires (1) a method for obtaining alternative sets A_x, (2) a metric with which to measure the distance of an utterance from its alternatives, and (3) a means with which to summarise distributions of pairwise distances.

Generating alternative sets. Since the 'true' alternative sets entertained by a human comprehender are not attainable, we generate them using neural autoregressive language models (LMs). For dialogue response generation, we use GPT-2 (Radford et al., 2019), DialoGPT (Zhang et al., 2020), and GPT-Neo (Black et al., 2021). For text generation, we use GPT-2, GPT-Neo, and OPT (Zhang et al., 2022). The text models are pre-trained, while dialogue models are fine-tuned on the respective

* This abstract presents work published under the title Information Value: Measuring Utterance Predictability as Distance from Plausible Alternatives in the *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5633–5653. Association for Computational Linguistics.

† Shared first authorship.

Proceedings of the Society for Computation in Linguistics (SCiL) 2024, pages 307-310.
Irvine, California, June 27-29, 2024

datasets. Further details on fine-tuning and perplexity scores are in (Giulianelli et al., 2023b). The resulting dataset, which contains 1.3M generations, is publicly available.[1]

Measuring distance from alternatives. We quantify the distance of a target utterance from an alternative production using three interpretable distance metrics, as defined by Giulianelli et al. (2023a). **Lexical:** Fraction of distinct n-grams in two utterances, with $n \in [1, 2, 3]$ (i.e., the number of distinct n-gram occurrences divided by the total number of n-grams in both utterances). **Syntactic:** Fraction of distinct part-of-speech (POS) n-grams in two utterances. **Semantic:** Cosine and euclidean distance between the sentence embeddings of two utterances (Reimers and Gurevych, 2019). These distance metrics characterise alternative sets at varying levels of abstraction (Katzir, 2007; Fox and Katzir, 2011; Buccola et al., 2022), enabling an exploration into the representational form of expectations over alternatives in human language processing.

Summarising distance distributions. Information value is a distribution over distances between an utterance y and the set of plausible alternatives (Equation 1). To summarise this distribution, we explore *mean* as the expected distance or the distance from a prototypical alternative, and *min* as the distance of y from the closest alternative production, implicating that proximity to a single alternative is sufficient to determine predictability.

Analysis

We study comprehension behaviour as recorded in contextualised acceptability judgements in text (CLASP; Bernardy et al., 2018) and in dialogue (SWITCHBOARD, DAILYDIALOG; Wallbridge et al., 2022) as well as eye-tracked (PROVO; Luke and Christianson, 2018) and self-paced reading times (BROWN; Smith and Levy, 2013).

Explaining psychometric data. Using information value, we investigate which dimensions of predictability effectively explain acceptability judgements and reading times. We also examine the effect of contextualisation on comprehension behaviour by defining two measures derived from information value, *context informativeness* and *out-of-context information value* (for definitions, see Giulianelli et al., 2023b). We evaluate each model relative to a baseline model which includes only control variables. As an indicator of predictive power, we report ΔLogLik, the difference in log-likelihood between a model and the baseline (Wilcox et al., 2020). We find that acceptability judgements factor in base rates of utterance acceptability (likely associated with grammaticality) but are predominantly driven by contextually modulated semantic expectations. In contrast, reading times are more influenced by the inherent plausibility of lexical items and part-of-speech sequences.

	SWITCHBOARD	DAILYDIALOG	PROVO
Surprisal	6.63	5.08	59.04
Information value			
Lexical	8.32	10.88	12.17
Syntactic	2.49	6.71	21.80
Semantic	34.20	30.41	6.86
All	43.11	35.42	45.19
Joint			
+ *Lexical*	14.08	10.23	72.60
+ *Syntactic*	9.77	8.05	75.70
+ *Semantic*	34.37	26.98	68.61
+ *All*	44.11	30.55	93.08

Table 1: ΔLogLik for surprisal, information value, and joint mixed effect models.

Relation to Surprisal. Focusing on acceptability judgements in the dialogue corpora and the reading times in PROVO, the psychometric measures for which we observed the highest explanatory power for information value, we fit linear mixed effect models with surprisal and information value in isolation and jointly as fixed effects. Table 1 summarises the results of this analysis. Information value is a stronger predictor of acceptability in written and spoken dialogue and is complementary to surprisal for predicting eye-tracked reading times.

Conclusion

Information value is a new way to measure predictability of utterances in terms of their distance from plausible continuations of the current linguistic context. It draws inspiration from the tradition of alternatives in semantics and pragmatics, and its psychometric predictive power is either higher or comparable, and almost always complementary, to that of aggregates of token-level surprisal. We hope information value will enable further investigation into the mechanisms involved in human utterance processing, and that it will serve as a basis for cognitively inspired learning rules and inference algorithms in computational models of language.

[1] *AltGen*: https://doi.org/10.5281/zenodo.10006413.

References

Suhas Arehalli, Brian Dillon, and Tal Linzen. 2022. Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 301–313, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kristijan Armeni, Roel M. Willems, and Stefan L. Frank. 2017. Probabilistic language models in cognitive neuroscience: Promises and pitfalls. *Neuroscience & Biobehavioral Reviews*, 83:579–588.

Jean-Philippe Bernardy, Shalom Lappin, and Jey Han Lau. 2018. The influence of context on sentence acceptability judgements. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461, Melbourne, Australia. Association for Computational Linguistics.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Brian Buccola, Manuel Križ, and Emmanuel Chemla. 2022. Conceptual alternatives: Competition in language and beyond. *Linguistics and Philosophy*, 45(2):265–291.

Danny Fox and Roni Katzir. 2011. On the characterization of alternatives. *Natural language semantics*, 19:87–107.

Richard Futrell and Roger Levy. 2017. Noisy-context surprisal as a human sentence processing cost model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 688–698, Valencia, Spain. Association for Computational Linguistics.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023a. What comes next? Evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14349–14371, Singapore. Association for Computational Linguistics.

Mario Giulianelli and Raquel Fernández. 2021. Analysing human strategies of information transmission as a function of discourse context. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 647–660, Online. Association for Computational Linguistics.

Mario Giulianelli, Sarenne Wallbridge, and Raquel Fernández. 2023b. Information value: Measuring utterance predictability as distance from plausible alternatives. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5633–5653, Singapore. Association for Computational Linguistics.

Roni Katzir. 2007. Structurally-defined alternatives. *Linguistics and philosophy*, 30:669–690.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, 41(5):1202–1241.

Steven G. Luke and Kiel Christianson. 2018. The Provo Corpus: A large eye-tracking corpus with predictability norms. *Behavior Research Methods*, 50(2):826–833.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally Typical Sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Claude E. Shannon. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Sarenne Wallbridge, Peter Bell, and Catherine Lai. 2022. Investigating perception of spoken dialogue acceptability through surprisal. In *Interspeech 2022: The 23rd Annual Conference of the International Speech Communication Association*, pages 4506–4510. International Speech Communication Association.

Jason Wei, Clara Meister, and Ryan Cotterell. 2021. A cognitive regularizer for language modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 5191–5202. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, pages 1707–1713. Cognitive Science Society.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.