Semantic Information: A difference that makes a difference

Anonymous ACL submission

Abstract

In this study based on an English fairytale corpus, we interpret Semantic information (SemI) in natural language as the difference of information between an informed and an uninformed system. Only an informed system contains SemI and its amount is the information difference between an informed and an uninformed system. This difference we were able to show.

1 Introduction

In this empirical study, based on an English fairytale corpus, we present an approach for measuring semantic information (SemI) in natural language, based on Shannon Information (Shannon, 1948) (SI) and inspired by the work of (Kolchinsky and Wolpert, 2018) who determine the following: SI is *syntactic information* (SynI) that quantifies statistical states in a system, or, additionally, between systems. In a linguistic context, we interpret the concept of a *system* as consisting of a linguistic production, a text or a corpus, and a language processor (LP).

An important difference betwen SynI and SemI is that the latter is *meaningful* information (Kolchinsky and Wolpert, 2018) which goes beyond a linguistic system: SemI draws on extralinguistic contexts and may include, for example, world knowledge, and only an informed system contains SemI. It is vital for the semantic interpretation, while an uninformed system lacks any meaningful information.

Building upon Kolchinsky and Wolpert (2018) framework, our approach hinges on the comparison between an informed LP and an uninformed one. We predict that *SemI will reduce the amount of surprisal a LP has to cope with*. Surprisal is a type of information and consequently an informed system, which is predicted to carry less information than an uninformed one. This may sound baffling, but our prediction means basically that facilitates text processing. We assume that SemI can be measured as the difference in surprisal between an informed and an uninformed system. The surprisal-difference between an uninformed and an informed system indicates the degree of SemI in the latter.

The proportionality relationship between processing effort and surprisal was established by (Hale, 2001). His *Surprisal theory* quantifies the predictability of a word as its surprisal, i.e., its negative logarithm of probability given the context.

In other words, we attribute *semantic surprisal* to an informed system, i.e., contextualised information (Tribus, 1961; Hale, 2001; Levy, 2008; Bentum, 2021) which we derive from semantic contexts.[1]. For the calculation of semantic surprisal, we employ a variant of the *Topic Context Model* (TCM) Kölbl et al. (2020, 2021); Philipp et al. (2022) (see Section 5). TCM outputs surprisal from a distribution of topics within a corpus, text or paragraph, and topics represent semantic concepts in the 'real' world. As surprisal in an uninformed system, we use *lexical surprisal*, derived from unigram probabilities of words. Lexical surprisal is SynI and thus equivalent with SI.

2 Points of departure and relevant work

In addition to the above-mentioned work by Kolchinsky and Wolpert (2018) and Hale (2001), further inspirations for our study are the works of Dretske (1981), and Floridi (2004, 2009). These approaches handle information-differences between two distinct systems and distinguish meaningful and meaningless information. Floridi (2004, 2009) terms this difference *strongly semantic information*. Inspired by these works are,

[1] Surprisal has been shown to be an empirically confirmed entity (e.g. DeLong et al. (2005); Bentum (2021)): it correlates with processes and states in the brain which makes it a concept of the *Philosophy of Mind*

among others, the studies of (Feldman and Peng, 2013; Peng et al., 2018; Rubino et al., 2016; Venhuizen et al., 2019) on idiom detection, translation-classification and predictive language comprehension, respectively. These studies have in common that information differences represent qualitative differences between a baseline condition and a special, surprising condition. For example, in (Feldman and Peng, 2013; Peng et al., 2018; Philipp et al., 2023a) the baseline condition includes sentences that can be understood literally, while the surprising, deviant condition comprises idiomatic sentences.

Different models for the calculation of surprisal have been employed, i.e., vector space representations of words (Feldman and Peng, 2013; Peng et al., 2018), n-gram models (Rubino et al., 2016), distributed situation-state space models (Venhuizen et al., 2019) and TCM (Kölbl et al., 2020, 2021; Philipp et al., 2022, 2023a,b) that, as mentioned above, is used in the present study.

3 Measuring semantic information

We start by defining a *system* as a language processor (LP) in an *environment* (a text/corpus/etc.). Following ideas by Hale (2001), we base our definition of SemI around the difficulty of the LP to process a text. The idea is thus: the more information (SemI) an LP has about a text, the less difficulty, i.e., SynI will be involved in parsing it. Another term for SynI in this context is *surprisal*; a term that suggests that the quantity at hand measures how surprising it is for the LP to encounter a word. In this view, SynI exists *within* an environment, while SemI is information *about* it. In particular, SemI is information about a text which helps *reduce* the surprisal that is measured while an LP parses the text. To be able to measure SemI thus requires us to compare two LPs: an informed and an uninformed one. We assess their respective processing difficulties by first considering the probability space of context-prediction pairs they can process. This could be, for example, bigrams where one word is known to the LP and the other one is its prediction based on that. On this space, we get three different distributions: the (un-)informed distributions by the two LPs, and the *actual* distribution directly taken from the text.

The processing difficulty can then be measured by comparing the LPs' distributions with the third one. In the practical part of this paper, we used the *Kullback Leibler Divergence* (KL) (Kullback and Leibler, 1951), but other types of metrics, such as perplexity, may also work. The interpretation of KL is that it measures how well a distribution P is approximated by another distribution Q. In terms of surprisal, it can be interpreted as the *average surprisal* an LP will experience working in a P-distributed environment while expecting a Q-distributed one. We denote it $KL(P, Q)$. Values yielded by that operation are always non-negative and reach 0 when they are identical. It is in general not symmetric. We thus get the two values $KL(T, U)$ and $KL(T, I)$ where U and I denote the uninformed and informed distributions respectively, and T denotes the 'text' or 'true' distribution. We can compare them using either Formula 1 or Formula 2.

$$\begin{aligned} SemI &= -log_2 \frac{KL(T,I)}{KL(T,U)} \\ &= \log_2 KL(T,U) - \log_2 KL(T,I) \end{aligned} \tag{1}$$

$$SemI = KL(T,U) - KL(T,I) \tag{2}$$

We deliberately give two formulae as possible alternatives because they measure slightly different things, each with the potential of having a unique merit. Both formulae measure how effectively semantic information has influenced the system. However, Formula 1 can be regarded as the relative version of Formula 2. Both can be interpreted giving the average reduction in the cost of text processing per word. However, in Formula 2 we measure the exact number of reduced bits from the average encoding length of a word, whereas in Formula 1, we consider the reduced amount of information relative to the average code lengths given the uninformed distribution.

In stark contrast to Shannon information, in the present setting there may be instances where the information content is negative. This happens when the informed LP experiences more processing difficulties than the uninformed one. In such cases, we may speak of 'disinformation' or 'deception', wherein a misleading expectation complicates the processing.

We emphasise that the specifics of how an LP operates are not rigidly defined. This is on purpose since different applications may call for different notions of uninformedness or different modes of operation of the LP.

4 Data

To test our prediction, we used an English fairytale corpus from INESC-ID Human Language Technology Lab[2](Lobo and De Matos, 2010). The corpus comprises 410 stories with in total 83,845 words. The average number of words per fairytale is 270. Preprocessing includes removing of all punctuation and converting them to lowercase, the words were already lemmatised. We split the texts into 300 training texts and 110 test texts.

5 Probability distributions and Workflow

5.1 The distributions

For every text, we need a total of three distributions: an *uninformed* one, an *informed* one, and the *actual* one. The uninformed distribution U has to be independent of the text, the informed one I has to depend on an informative token extracted from the text, and the actual one T is the real distribution of words in the text.

For the uninformed distribution, we choose for the probability function the relative frequency of every word in the training corpus. Before normalising however, we add 10^{-17} to all words, including those that do not make an appearance in the training corpus, so as to prevent a division by 0 when the KL-divergence is computed. Hence, the distribution is given by Formula 3.

$$P_U(w) = \frac{N + 10^{-17}}{\sum_{w \in \text{training and test corpus}} (N + 10^{-17})} \quad (3)$$

Where N is the number of occurrences of w in the training corpus. For the informed distribution, we make use to the *Topic Context Model* (TCM) (Kölbl et al., 2020, 2021; Philipp et al., 2022, 2023a,b)[3]. The TCM is an extended topic model, calculating the probability of a word w given a distribution of *topics* for the text or corpus the word appears in.

In this study, we employ the TCM based on *Latent Dirichlet Allocation* (Blei et al., 2003) (LDA). We initialise LDA with $n = 10$ topics and train it on the training corpus. This gives us for each topic a probability distribution $P(w_i|t_i)$ that indicates the probability a word is associated to a specific topic. We can define the *topic space* as the simplex $\{(x_1, x_2, \ldots, x_n) \in [0,1]^n | \sum x_k = 1\}$. Then for each document d, its *topic vector* v_d is an element of the topic space whose coordinates are given by the probabilities $P(t_i|d)$ that any given word in d is associated to topic t_i. Now the informed distribution for a word w given the topic vector v_d of a document is given by Formula 4.

$$P_I(w|v_d) = \sum_{i=1}^{n} P(w|t_i)P(t_i|d) \quad (4)$$

5.2 Workflow

We compute P_U once at the beginning and then we compute for every document d in the test set four probability functions: P_T, $P_I^{(i)}$, $P_I^{(ii)}$, and $P_I^{(iii)}$. Here, P_T is the probability function of T. The other three are three different informed distributions, each computed with a different topic vector: $P_I^{(i)}$ uses v_d, i.e., the correct topic vector; $P_I^{(ii)}$ uses v_0, i.e., the topic vector of the first document in the test set; $P_I^{(iii)}$ uses a randomly generated element of the topic space. Then we calculate $KL(T, U)$ and the three different versions of $KL(T, I)$. From these we calculate for each $KL(T, I)$ the pair of SemI measures given in Formulas 1 and 2.

6 Results

Figure 1 displays the values of the test set as calculated via Formula 1. The left-most plot indicates the distribution of the SemI values where the correct topic vector was used for every document. The middle plot indicates the values where one of the topic vectors was fixed for every document. Lastly, the right-most plot indicates the distribution in the case of randomly generated topic vectors. We call these three cases (i), (ii), and (iii), corresponding to the probability functions $P_I^{(i)}$, $P_I^{(ii)}$, and $P_I^{(iii)}$ defined in Section 5.1 It can be seen that correct topic vectors carry the highest amount of semantic information and the randomly generated topic vectors yield negative values; this indicates that the LP is actively confused by the hint; the average surprisal of every word grows. Interestingly the mismatched topics still yield relatively high values, albeit to a lesser extent.

The situation in Figure 2 is analogous but the values were generated with Formula 2.

In all three cases, the difference in surprisal values between the informed and systems is sig-

[2] https://www.hlt.inesc-id.pt/w/Fairy_tale_corpus

[3] https://github.com/jnphilipp/tcm

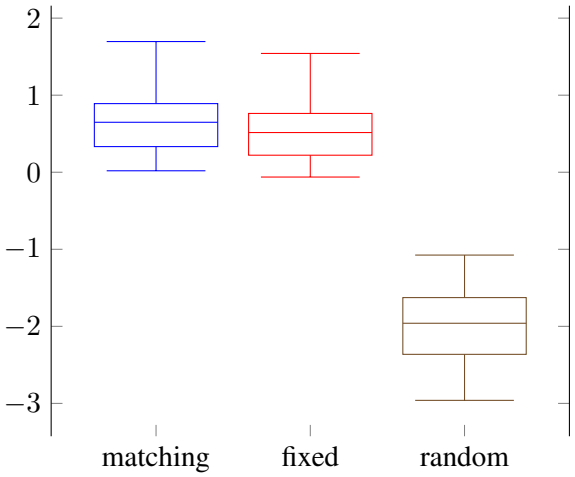

Figure 1: SemI calculated with Formula 1. The amount of SemI can be read off the y-axis.

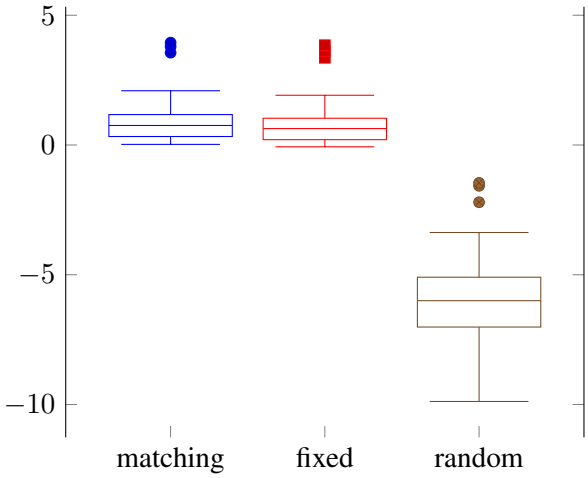

Figure 2: SemI calculated with Formula 2. The amount of SemI can be read off the y-axis.

nificant (case (i): $t = -8.9, p \approx 0$; case (ii): $t = -7.6, p \approx 0$; case (iii): $t = 43.1, p \approx 0$).

7 Conclusion and future work

This pilot study yielded the predicted results and thus provided clues to the substance of the prediction: an LP that is supplied with a token carrying semantic information performs better than one that is not. A reduction of surprisal is evident for matching and fixed topic vectors as contexts, whereas an increase can be observed with random topic vectors. For the fixed topic vectors, the reduction in each document is smaller than that for matching topic vectors, but it still occurs. The first part can be explained by the fact that fairytales are similar enough to still supply *some* semantic information, but not as much as a precise topic vector ever could. One could say that the LP was made aware of the fact that it is processing a fairytale, but not which one. Any future work should include comparisons between genres to see if this effect does in fact become stronger with less closely related texts. In the case of random topic vectors, we can see that wrong expectations lead to confusion on the part of the LP.

However, it should be noted that the connection between surprisal and semantics is not straightforward. The reduction of surprisal can only give an indirect indication of semantics: for text processing, which is always also about meaning, semantic surprisal ensures a lower processing effort, that is, the LP has to process not so much information. SemI in our interpretation represents the amount of higher certainty in language processing. The assumption that the difference in surprisal between informed and uninformed systems has a semantic quality is plausible, since this difference is not due to purely structural semantic differences in the system, i.e., texts, which is why the difference is also not a syntactic but a semantic one.

In this study, we restricted ourselves to computing the SemI values of given informing (or disinforming) tokens. However, the results indicate this method's potential for applications knowledge extraction: among a set of tokens, the one with the highest semantic information may reveal useful knowledge about the underlying text. Also, there may be many different types of informative token besides topic vectors, such as keywords or text genre.

Moreover, our concept of semantic information implies that knowledge about a system would have to be taken into account, which can be subsumed under the term *world knowledge*. These are desiderata of future research.

Limitations

The fairytale corpus is quite small; future studies would have to be based on larger corpora. The same goes for the literary genre, future corpora would need to use different genres. Moreover, our concept of semantic information implies that advanced knowledge about a system, i.e., knowledge of the world, would have to be taken into account. After all, our pilot study does not provide yet the data basis for incorporating our findings into practical applications, such as the automatic detection of disinformation.

References

Martijn Bentum. 2021. *Listening with great expectations: A study of predictive natural speech processing*. Ph.D. thesis, [Sl]:[Sn].

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Katherine A DeLong, Thomas P Urbach, and Marta Kutas. 2005. Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, 8(8):1117–1121.

Fred Dretske. 1981. *Knowledge and the Flow of Information*. MIT Press.

Anna Feldman and Jing Peng. 2013. Automatic detection of idiomatic clauses. In *Computational Linguistics and Intelligent Text Processing: 14th International Conference, CICLing 2013, Samos, Greece, March 24-30, 2013, Proceedings, Part I 14*, pages 435–446. Springer.

Luciano Floridi. 2004. Outline of a theory of strongly semantic information. *Minds and machines*, 14:197–221.

Luciano Floridi. 2009. Philosophical conceptions of information. In *Formal theories of information: From Shannon to semantic information theory and general concepts of information*, pages 13–53. Springer.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

Artemy Kolchinsky and David H Wolpert. 2018. Semantic information, autonomous agency and non-equilibrium statistical physics. *Interface focus*, 8(6):20180041.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Max Kölbl, Yuki Kyogoku, J. Nathanael Philipp, Michael Richter, Clemens Rietdorf, and Tariq Yousef. 2020. Keyword Extraction in German: Information-theory vs. Deep Learning. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 1: NLPinAI*, pages 459–464. INSTICC, SciTePress.

Max Kölbl, Yuki Kyogoku, J. Nathanael Philipp, Michael Richter, Clemens Rietdorf, and Tariq Yousef. 2021. The semantic level of shannon information: Are highly informative words good keywords? a study on german. In Roussanka Loukanova, editor, *Natural Language Processing in Artificial Intelligence - NLPinAI 2020*, volume 939 of *Studies in Computational Intelligence (SCI)*, pages 139–161. Springer International Publishing.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Paula Vaz Lobo and David Martins De Matos. 2010. Fairy tale corpus organization using latent semantic mapping and an item-to-item top-n recommendation algorithm. In *LREC*, volume 10, pages 1472–1475.

Jing Peng, Anna Feldman, and Ekaterina Vylomova. 2018. Classifying idiomatic and literal expressions using topic models and intensity of emotions. *arXiv preprint arXiv:1802.09961*.

J Nathanael Philipp, Max Kölbl, Erik Daas, Yuki Kyogoku, and Michael Richter. 2023a. Perplexed by idioms? In *Knowledge Graphs: Semantics, Machine Learning, and Languages*, pages 70–76. IOS Press.

J. Nathanael Philipp, Max Kölbl, Yuki Kyogoku, Tariq Yousef, and Michael Richter. 2022. One step beyond: Keyword extraction in german utilising surprisal from topic contexts. In *Intelligent Computing*, pages 774–786, Cham. Springer International Publishing.

J. Nathanael Philipp, Michael Richter, Erik Daas, and Max Kölbl. 2023b. Are idioms surprising? *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*.

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 960–970, San Diego, California. Association for Computational Linguistics.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Myron Tribus. 1961. Information theory as the basis for thermostatics and thermodynamics.

Noortje J Venhuizen, Matthew W Crocker, and Harm Brouwer. 2019. Semantic entropy in language comprehension. *Entropy*, 21(12):1159.