Sheaves and garden-path sentences

Daphne Wang Mehrnoosh Sadrzadeh

University College London

The most prominent way of predicting human reading times in computational linguistics is by using the information-theoretic surprisal[1, 4], defined as:

$$S = -\log_2 P\left[w_i | w_1 \dots w_{i-1}\right]$$

And indeed, a strong correlation has been observed between surprisal and human reading times in the general case[1]. However, the limitations of this approach have been observed, for example in garden-path sentences[3]; i.e. sentences that are grammatically correct, but lead the reader to commit to an ultimately incorrect parse.

Sheaves are mathematical objects that have been used in algebraic topology and logic to describe the passage from local to global data (e.g. functions on continuous spaces)[2]. The data of a sheaf is essentially a map F from a domain space to a base space. Each element of the base space (e.g. intervals) is then associated with a set of *local data* in the domain space, which corresponds to all elements in the domain which are mapped to it. The elements of the local data are then called *sections*, and the collection of local data is called a *presheaf*. A *sheaf* is then a presheaf which agrees on intersections on the base space; i.e. its data is *globally consistent*.

In this work, we use sheaf theory to model an incremental (parallel-ranked) model of parsing. We then used a quantitative measure of "sheafness" of statistical data obtained from the transformer-based language model `BERT` to measure the difficulty of reading a given word, and use these measures to successfully predict human reading times.

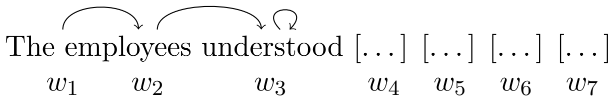
In our approach, we started by modelling the incremental increase in linguistic context as a sequence of subphrases of a given sentence of increasing length. For example, given the sentence "The employees understood the contract would change", we could have the set of contexts as:

$$\{\textit{The}, \textit{The employees}, \textit{The employees understood}, \textit{The employees understood the}, \dots\}$$

This will be our base space. These contexts can then be ordered by the prefix order $\leq$, i.e. we have for example $\textit{The employees} \leq \textit{The employees understood}$. We then encoded the process of choosing the grammatical structure with a phrase as a function from each of the words in the sentence to their head in a dependency grammar. Formally, this can be seen as taking a collection of sections on the presheaf $\mathcal{E}$ where:

$$\mathcal{E}(c) = \{f : \{w_1, \dots, w_k\} \to \{w_1, \dots, w_N\}\}$$

for any context c, with N being the length of the sentence, k being the length of the context c, and $\{w_1, \dots, w_N\}$ being the list of words in the sentence. For example the parse:


The employees understood [...] [...] [...] [...]
w_1 w_2 w_3 w_4 w_5 w_6 w_7

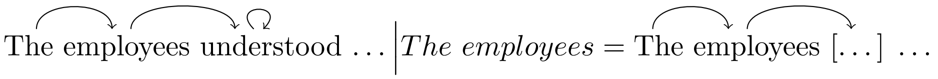
of the context *The employees understood* can be seen as the function:

$$f = [w_1 \mapsto w_2, w_2 \mapsto w_3, w_3 \mapsto w_3]$$

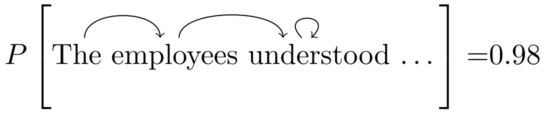
We can also define restriction maps as:

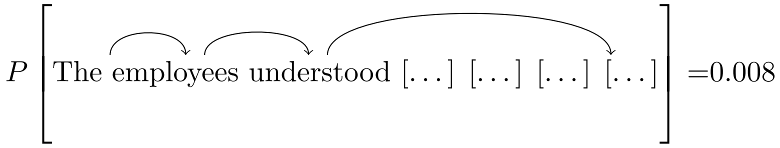
$$\begin{aligned} & f_{c_1} = [w_1 \mapsto n_1, \dots, w_k \mapsto n_k, w_{k+1} \mapsto n_{k+1}, \dots w_m \mapsto n_m] \\ \implies & f_{c_1}|c_2 = [w_1 \mapsto n_1, \dots, w_k \mapsto n_k] \end{aligned}$$

whenever $c_1 \leq c_2$ ($f_{c_1}|c_2$ is understood as f_{c_1} restricted to the context c_2). So, for example:


The employees understood ...
The employees = The employees [...] ...

Now, in the general case, there is not a unique possible choice of parse of a given subphrase (i.e. most subphrases of a sentence are syntactically ambiguous); and each of these different possible parses comes with a degree of likelihood. For example, we could have:





$$\vdots$$

This can therefore be modelled by taking sections over a new presheaf D which associates all possible probability distributions over parses, given a context c. Given a family $d = \{d_c\}_{c \in \mathcal{C}}$ of probability distributions over contexts in $\mathcal{C}$ (i.e. a collection of sections of D), we then say d is *compatible* whenever:

$$d_c|c \cap c' = d_{c'}|c \cap c'$$

for any choice of contexts c, c'. We then define our degree of inconsistency SF of d to be the minimum value such that:

$$d = (1 - \mathsf{SF}) \cdot d_{comp} + \mathsf{SF} \cdot d'$$

where multiplication is done point-wise, d_{comp} is a compatible set of sections, and d' can be any (not necessarily compatible) set of sections. This quantity, SF, measures the reader will have to change their mental representation of the grammatical structure of the sentence upon encoutering new information. In order to model an incremental reading process, we consider pairs of contexts which only differ by a word (e.g. *The employees* and *The employees understood*), and hypothesise that SF correlates with the reading time of the extra word (e.g. *understood*).

Here, we work from the dataset of [5], and computed probability distributions from predictions from the language model `BERT`, and the respective dependency parses obtained from `spaCy`. We also compared various `BERT` language models (`dilstilbert`, `bert-base-cased` and `bert-large-cased`), as well as different models of `spaCy` (`en_web_core_sm`, `en_web_core_lg`, `en_web_core_trf`). We then calculate the correlations between our degree of inconsistency SF and the reading times obtained from [5] and observed that there is indeed a strong linear correlation between them; the Pearon's ρ and associated p-values are quoted in Table 1. We then use a linear regression model to predict reading times and compared it with the existing literature[3]. We found that our measure produced more accurate predictions than the ones from surprisal, see Table 2, although predicted garden-path effects are still underestimated. In addition, our consistency measure is able to predict statistically significant differences between hard NP/Z and easy NP/S garden-path sentences, see Table 3, which was previously impossible to do using surprisal[3].

We believe our approach provides a straightforward way of creating mathematical models of psycholinguistic theories, and can evaluate these theories empirically. We also aim to investigate other factors of human comprehension using this framework as future work; for example, comparing repair-based approaches as opposed to reanalysis, including the influence of plausibility and investigating the interplay between surprisal and SF seem to be promising avenues of research.

		BERT model		
		distilbert	bert-base	bert-large
spaCy model	en_core_web_sm	0.64 (0.008)	0.80 (0.0002)	0.79 (0.0003)
	en_core_web_lg	0.63 (0.009)	0.79 (0.0003)	0.78 (0.0003)
	en_core_web_trf	0.67 (0.004)	0.78 (0.0004)	0.76 (0.0006)

Table 1: Pearson's ρ coefficients and associated p-values (in brackets) between SF and reading times.

	Prediction (ms)		**Observed** (ms)
	SF	S[3]	
NP/S	**63**	24	87
NP/Z	**110**	30	400

Table 2: Garden-path effect predictions. The BERT model used is bert-base-cased and the spaCy model used (for SF calculations) is web_core_trf.

		BERT model		
		distilbert	bert-base	bert-large
spaCy model	en_core_web_sm	0.03	0.01	0.09
	en_core_web_lg	0.02	0.04	0.24
	en_core_web_trf	0.39	0.0001	0.01

Table 3: p-values associated with the t-test evaluating whether the garden-path effect predictions obtained from SF for NP/S and NP/Z are sampled from the same distribution.

References

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

[2] S. MacLane and I. Moerdijk. *Sheaves in geometry and logic: A first introduction to topos theory.* Springer Science & Business Media, 2012.

[3] M. Van Schijndel and T. Linzen. Modeling garden path effects without explicit hierarchical syntax. In *CogSci*, 2018.

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

[5] P. Sturt, M. Pickering, and M. Crocker. Structural change and reanalysis difficulty in language comprehension. *Journal of Memory and Language*, 40:136–150, 1999.