DisCoCirc

Vincent Wang-Maścianica
Quantum Group
Department of Computer Science
The University of Oxford
vincent.wang@cs.ox.ac.uk

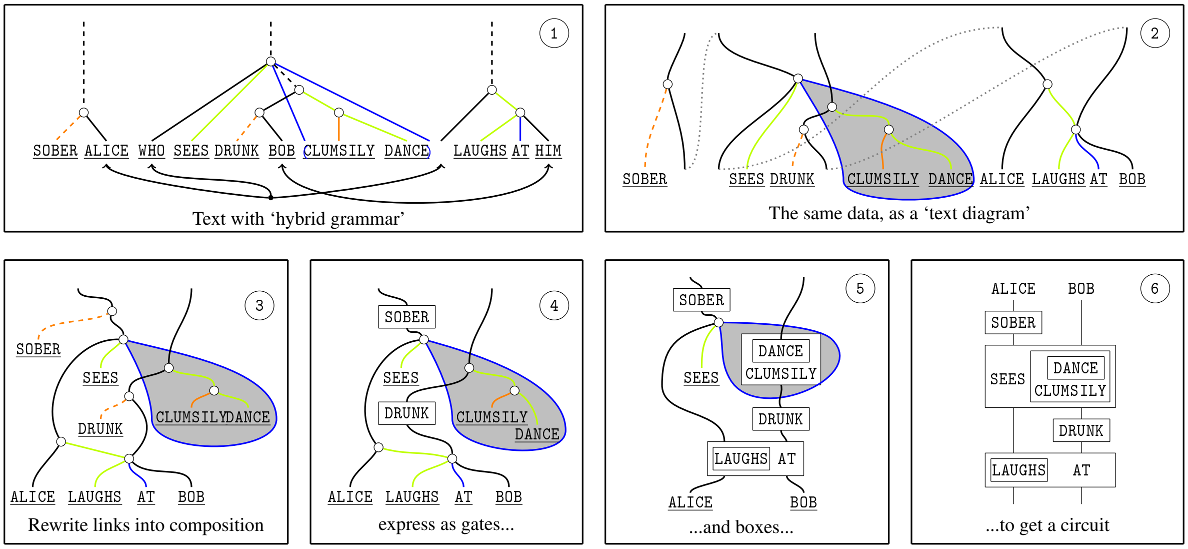
Jonathon Liu
jonathon.liu@quantinuum.com

Bob Coecke
bob.coecke@quantinuum.com

Compositional Intelligence
Quantinuum
Oxford

DisCoCirc was proposed in *The Mathematics of Text Structure* [1] as a framework for language that is distributional (a.k.a. vectorial), and compositional in the same manner as circuits (i.e. process-theoretically, as in symmetric monoidal categories). It is a lean structure which aims to capture fundamental linguistic 'connections' between meanings (e.g. as described by grammar). Further, it can represent general text beyond just sentences and is therefore a significant evolution of the existing DisCoCat framework [2]. Indeed, we believe that the 2-dimensional representation afforded by string diagrams is a more natural habitat for language than a 1-dimensional string of symbols, and that this passage from 1D to 2D eliminates much of the extraneous bureaucracy associated with grammar and other stylistic conventions.

Recently, this framework was formalized for a substantial fragment of English in *Distilling Text into Circuits* [4]. In this paper, a simple 'hybrid' model of grammar and discourse semantics was introduced for the purpose of showing that all text in this fragment of English can be represented as a circuit.



Text with 'hybrid grammar'

The same data, as a 'text diagram'

Rewrite links into composition

express as gates...

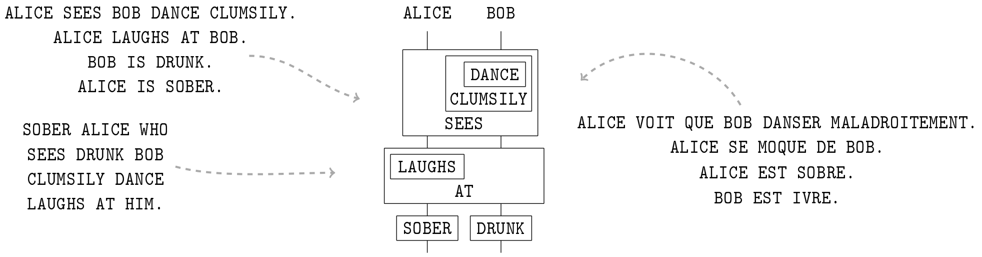
...and boxes...

...to get a circuit

The hybrid grammar incorporated elements of generative context-free grammars, as well as features describing coreference, pronouns, and phrase scope. Furthermore, it was shown that this map is surjective – i.e. that any possible text circuit as presented in the paper is realizable as an English text equipped with the hybrid grammar.

One interesting aspect of DisCoCirc is that it displays a significant degree of intra- and inter-language independence, where different texts within and across languages conveying the same content have the same circuit representation. These language independence properties have since been examined in *Grammar Equations* [3] and *Language independence of DisCoCirc's Text Circuits: English and Urdu* [5].

Submitted to:
ACT 2023



In view of natural language processing, a major practical advantage of DisCoCirc over existing semantic formalisms is that it naturally accommodates the power of machine learning. Since text diagrams are just formal string diagrams, one is free to functorially endow them with semantics in various ways. For instance, one can use various kinds of vector embeddings, neural networks, or higher linear-algebraic operations. In particular, given the representation of text as circuits, DisCoCirc is uniquely well-suited to quantum natural language processing. In this case, one can convert DisCoCirc structures into a quantum circuit by filling in the boxes with parametrized ansatzes.

There is significant ongoing practical work involving DisCoCirc. Firstly, an automated 'parsing pipeline' that converts English text to its DisCoCirc representation is being developed. The pipeline achieves good coverage over real-world text, and in particular it covers the fragment of English discussed in [4]. It works by converting syntax trees from categorial grammar into expressions from simply-typed lambda calculus, and then performing transformations on these lambda expressions, before finally interpreting the expressions string-diagrammatically. The outputs of this pipeline are being applied to the bAbI QA tasks [6], with promising results. These are a set of simple question-and-answer NLP tasks serving as a first proof of concept. The aim is to learn meaningful and interpretable representations and show that the semantics of natural language can be understood as a compositional process with a clear flow of information. In particular, DisCoCirc-based models have the advantage over contemporary state-of-the-art black box NLP models of being intrinsically interpretable and structured. Approaches involving both 'filling in the boxes' with classical neural nets, as well as filling them in with parametrized quantum circuits, are being explored.

References

[1] B. Coecke (2021): *The Mathematics of Text Structure*, pp. 181–217. Springer International Publishing. ArXiv:1904.03478.

[2] B. Coecke, M. Sadrzadeh & S. Clark (2010): *Mathematical foundations for a compositional distributional model of meaning*. In J. van Benthem, M. Moortgat & W. Buszkowski, editors: *A Festschrift for Jim Lambek*, *Linguistic Analysis* 36, pp. 345–384. Arxiv:1003.4394.

[3] B. Coecke & V. Wang (2021): *Grammar Equations*. *arXiv preprint arXiv:2106.07485*.

[4] Vincent Wang-Mascianica, Jonathon Liu & Bob Coecke (2023): *Distilling Text into Circuits*. Available at `http://arxiv.org/abs/2301.10595`. ArXiv:2301.10595 [cs, math].

[5] M. H. Waseem, J. Liu, V. Wang-Maścianica & B. Coecke (2022): *Language-independence of DisCoCirc's Text Circuits: English and Urdu*. In M. Moortgat & G. Wijnholds, editors: *Proceedings End-to-End Compositional Models of Vector-Based Semantics*, *Electronic Proceedings in Theoretical Computer Science* 366, Open Publishing Association, pp. 50–60, doi:10.4204/EPTCS.366.7.

[6] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin & Tomas Mikolov (2015): *Towards ai-complete question answering: A set of prerequisite toy tasks*. *arXiv preprint arXiv:1502.05698*.