Categorical Models of Discourse

Lachlan McPheat
supervised by Mehrnoosh Sadrzadeh
University College London

The Lambek calculus as a model of natural language provides the desirable structural property of compositionality which complements the distributionality of statistical models of language. These two properties were fused using monoidal biclosed categories in [1]. Applications of distributional-compositional models to discourse analysis have remained theoretical, and vector space interpretations have not been experimented with. The main hurdle to modelling discourse using existing models is that the phenomena governing the composition of sentences to form a discourse differ from those composing words to form sentences, i.e. grammar. An important inter-sentential phenomenon for forming discourses is coreference, which occurs when two or more lexically distinct words or phrases have identical semantics. A simple example is pronominal anaphora such as in the discourse *Sam runs. They trip.* where *They* clearly has the same meaning as *Sam*. Adding structural rules to Lambek calculus in the form of relevant modalities [3] allowed for parsing parasitic gaps, a complex movement phenomenon, and coreference. However, in the same paper the resulting logic was proven undecidable. A more recent extension of Lambek calculus using soft subexponentials [2] has a decidable fragment which has sufficient expressive power for parsing coreference and parasitic gaps. We develop categorical semantics for these logics in terms of monoidal biclosed categories with added structures, forming categorical models of discourse. We define functors into the category of finite dimensional real vector spaces, which preserve these structures [4]. The categorical semantics let us parse discourses using string diagrams, which in turn are interpreted as meaning in the vector space semantics, and has been experimented with on NLP tasks such as phrase similarity and disambiguation.

References

[1] B. Coecke, M. Sadrzadeh, and S. Clark. Mathematical Foundations for Distributed Compositional Model of Meaning. Lambek Festschrift. *Linguistic Analysis*, 36:345–384, 2010.

[2] M. Kanovich, S. Kuznetsov, V. Nigam, and A. Scedrov. Soft Subexponentials and Multiplexing. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pages 500–517. 2020.

[3] M. Kanovich, S. Kuznetsov, and A. Scedrov. Undecidability of the Lambek calculus with a relevant modality. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9804 LNCS:240–256, 2016.

[4] L. McPheat, M. Sadrzadeh, H. Wazni, and G. Wijnholds. Categorical vector space semantics for lambek calculus with a relevant modality,https://arxiv.org/abs/2005.03074, 2020.