Mathematical Foundations for a Compositional Distributional Model of Meaning (Coecke et al., [2010])

Overview.

In this paper, the authors develop an NLP model from a category theoretic construction of language. The major insight of the paper is that we can infer the meaning of a sentence from the meanings of the subsentences and words that compose it.

In order to model sentences, the authors use a product category to separate the composition of meaning from the composition of structure. This allows the model to incorporate both distributional (learned from data) and symbolic (rule based) components. The symbolic structure enforces grammatical correctness, and the distributional structure embodies the meanings of words and subsentences.

The authors use a standard distributional approach by representing words as vectors in a vector space. They choose to use the tensor product to represent the composition of multiple words, which allows them to enforce compact closure and avoid losing information with vector composition. Importantly, the authors choose to use vectors to represent "atomic" words like nouns but use linear maps to represent verbs. This allows them to model the role of a verb applying an "action" to a noun or pair of nouns. They do not provide much detail about the contents of the vectors themselves (word2vec, TF-IDF, etc) and instead leave this to future work.

To model the algebra of grammar, the authors choose to use a free pregroup generated by primitive language constructs. The pregroup morphisms represent the composition of words and subsentences. This approach allows them to model the grammmatical roles of words, subsentences of varying lengths and full sentences as elements of the same algebraic structure. Furthermore, since pregroups are compact closed categories, the product category of meaning vectors and grammatical structure is compact closed as well.

The authors also indicate how we can use this system to evaluate the truth of a positive transitive sentence by using the verb map to project the sentence representation from the space spanned by the noun basis onto a one dimensional space spanned by the basis vectors 1 (true) and 0 (false).

Comments.

First, the authors' choice of tensor product to represent word vector composition seems likely to present computational challenges for long sentences, since the vector space dimensionality explodes after a few composed tensor products. Furthermore, the decision to represent atomic words (like nouns) as vectors and words with compound types (like verbs) as linear maps is a little strange. On one hand, this provides us with a computational mechanism to express words acting on each other. On the other hand, this seems to challenge and limit the pregroup's role as the representation of a generic grammatical structure.

REFERENCES

[1] Coecke, B., Sadrzadeh, M., and Clark, S. (2010). Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*.