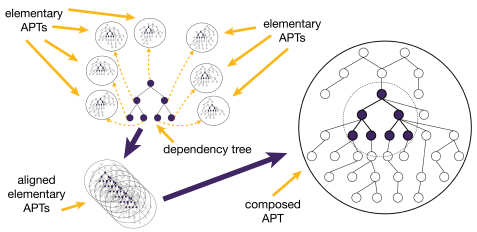
Improving Sparse Word Representations with Distributional Inference for Semantic Composition

Thomas Kober, Julie Weeds, Jeremy Reffin and David Weir
TAG Laboratory, University of Sussex

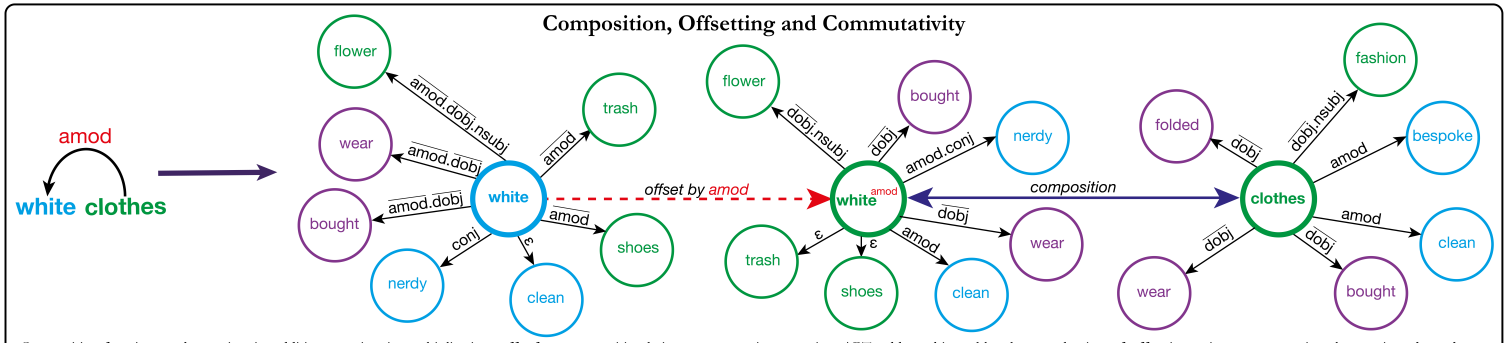
Anchored Packed Dependency Trees (APTs) - A Theory of Composition for Distributional Semantics

Distributional Composition with APTs



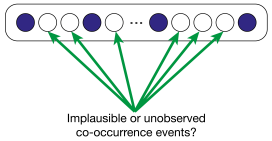
APTs are based on dependency trees and represent phrases and full sentences in the same space as single lexemes. Composition is a process of contextualising the lexemes being composed such that each lexeme is bespoke to that particular context (Weir et al., *CL special isse on Formal Distributional Semantics*, 2016).

Composition, Offsetting and Commutativity



Composition functions such as pointwise addition or pointwise multiplication suffer from composition being a commutative operation. APTs address this problem by a mechanism of offsetting a given representation along a given dependency relation when composing two lexemes. This process aligns the feature spaces of two lexemes with different parts of speech by leveraging the information from their dependency relations. For example when composing the adjective white with the noun clothes, one would offset the adjective by the amod in order to project it into noun space. This process makes composition a non-commutative operation.

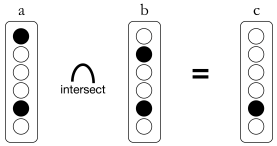
The Data Sparsity Challenge



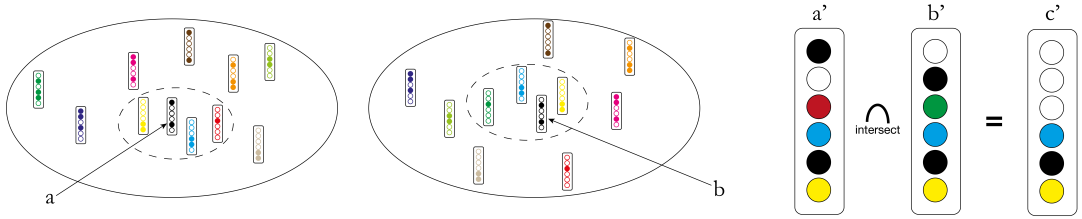
Data sparsity causes incomplete representations due to not observing a large number of plausible co-occurrences. Part of the problem is estimating whether an unobserved event is implausible or due to sparsity.

Data sparsity is especially problematic for intersective composition functions that aim to disambiguate the meaning of a phrase in context.

It is not feasible to apply common dimensionality reduction techniques because the structured nature of the APT space needs to be preserved for composition and they primarily address the model sparsity problem.



Distributional Inference (DI)



Exploiting the distributional neighbourhood by explicitly inferring unobserved but plausible co-occurrences enriches elementary word representations. This in turn results in richer composed representations. Using an intersective composition function eliminates co-occurrences unrelated to the phrase as a whole, while keeping plausible features.