Incremental Distributional Semantics for Dynamic Syntax

Mehrnoosh Sadrzadeh, Ruth Kempson, Matt Purver
Queen Mary University of London

Distributional semantics is inspired by ideas of Firth and Harris, the former of which said 'you shall know a word by the company it keeps'[Fir57]. Natural Language Processing researchers used this idea to turn corpora of documents into co-occurrence matrices to record frequencies of co-occurrences between words. These matrices are vector spaces wherein target words are vectors. The distances between these vectors represent their semantic similarity[RG65, SWY75].

Distributional semantics does not naturally model phrases and sentences: these constructions are compositional, take a variety of different forms, which often do not repeat in a corpus. Compositional distributional semantics (CDS) remedies this problem[CSC10, KM13]. CDS is guided by the grammatical structures of sentences, usually given in a type-logical system such as CCG, Lambek Calculus, or Pregroup Grammar. The CDS's work on complete parse trees of sentences and do not take into account the incremental behaviour of language, as observed in dialogue interactions and phenomena such as ellipsis. These are indeed accounted for in the grammatical formalism of Dynamic Syntax (DS) [RK15].

In this paper, we take a first step towards remedying this problem. From a high level stance, a CDS is a structure preserving map $\mathcal{F}$ from grammatical structures to vector representations:

$$\mathcal{F}\colon \text{Grammar} \implies \text{Vectors Semantics}$$

The input to $\mathcal{F}$, is a fully parsed grammatical structure. In order to take into account incrementality, we need to work with sequences of partial structures. We extend $\mathcal{F}$ to the following $\mathcal{G}$ map:

$$\mathcal{G}\colon (\Sigma \times \text{DST})^n \implies (\text{Vectors Semantics})^n$$

The input to $\mathcal{G}$ is a a sequence of word-tree $(w, \mathtt{T})$ pairs, for $w \in \Sigma$ and $\mathtt{T}$ a DS tree. Its output is a sequence of vectors, each constructed inductively from the trees and actions of DS. $\mathcal{G}$ mimics the ongoing build-up process of DS within the vector models: it takes transitions as the core of the explanation, rather than as operations defined on the output. The results obtained so far shows that the setting provides a means of formulating an incremental notion of content for emergent construal of ellipsis without anything other than the vectorial notion of content.

Our setting allows for composition and incrementality, with composition happening within sentences and incrementality between (sometimes within) them. The incrementality that we have modelled so far is guided by grammar. Since CDS is based on a type-logic, we first need to parse the sentences, then transform them. Thus certain substitution actions have to wait for the sentences to be completed to take effect. This is not be a full representative of the actions that happens in mind. We are working on this issue for the paper version of the abstract.

References

[CSC10] B. Coecke, M. Sadrzadeh, and S. Clark. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36:345–384, 2010.

[Fir57] J.R. Firth. A synopsis of linguistic theory 1930–1955. In *Studies in Linguistic Analysis*. 1957.

[KM13] Jayant Krishnamurthy and Tom M. Mitchell. Vector space semantic parsing: A framework for compositional vector space models. In *Proceedings of the 2013 ACL Workshop on Continuous Vector Space Models and their Compositionality*, 2013.

[RG65] H. Rubenstein and J.B. Goodenough. Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10):627–633, 1965.

[RK15] E. Gregoromichelaki M. Purver R. Kempson, R. Cann. Ellipsis. In *Handbook of Contemporary Semantic Theory (2nd editn)*, pages 156–194. Blackwell, 2015.

[SWY75] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18:613–620, 1975.