Improving Semantic Composition with Offset Inference

Thomas Kober, Julie Weeds, Jeremy Reffin and David Weir
TAG laboratory, Department of Informatics, University of Sussex
Brighton, BN1 9RH, UK
{t.kober, j.e.weeds, j.p.reffin, d.j.weir}@sussex.ac.uk

Abstract
Count-based distributional semantic models suffer from sparsity due to unobserved but plausible co-occurrences in any text collection. This problem is amplified for models like Anchored Packed Trees (APTs), that take the grammatical type of a co-occurrence into account. We therefore introduce a novel form of distributional inference that exploits the rich type structure in APTs and infers missing data by the same mechanism that is used for semantic composition.

1 Introduction
Anchored Packed Trees (APTs) is a recently proposed approach to distributional semantics that takes distributional composition to be a process of lexeme contextualisation (Weir et al., 2016). A lexeme’s meaning, characterised as knowledge concerning co-occurrences involving that lexeme, is represented with a higher-order dependency-typed structure (the APT) where paths associated with higher-order dependencies connect vertices associated with weighted lexeme multisets. The central innovation in the compositional theory is that the APT’s type structure enables the precise alignment of the semantic representation of each of the lexemes being composed. Like other count-based distributional spaces, however, it is prone to considerable data sparsity, caused by not observing all plausible co-occurrences in the given data. Recently, Kober et al. (2016) introduced a simple unsupervised algorithm to infer missing co-occurrence information by leveraging the distributional neighbourhood and ease the sparsity effect in count-based models.

In this paper, we generalise distributional inference (DI) in APTs and show how precisely the same mechanism that was introduced to support distributional composition, namely “offsetting” APT representations, gives rise to a novel form of distributional inference, allowing us to infer co-occurrences from neighbours of these representations. For example, by transforming a representation of white to a representation of “things that can be white”, inference of unobserved, but plausible, co-occurrences can be based on finding near neighbours (which will be nouns) of the “things that can be white” structure. This furthermore exposes an interesting connection between distributional inference and distributional composition. Our method is unsupervised and maintains the intrinsic interpretability of APTs.

2 Offset Representations
The basis of how composition is modelled in the APT framework is the way that the co-occurrences are structured. In characterising the distributional semantics of some lexeme \( w \), rather than just recording a co-occurrence between \( w \) and \( w' \) within some context window, we follow Padó and Lapata (2007) and record the dependency path from \( w \) to \( w' \). This syntagmatic structure makes it possible to appropriately offset the semantic representations of each of the lexemes being composed in some phrase. For example many nouns will have distributional features starting with the type amod, which cannot be observed for adjectives or verbs. Thus, when composing the adjective white with the noun clothes, the feature spaces of the two lexemes need to be aligned first. This can be achieved by offsetting one of the constituents, which we will explain in more detail in this section.

We will make use of the following nota-
tion throughout this work. A typed distributional feature consists of a path and a lexeme such as in \texttt{amod:white}. Inverse paths are denoted by a horizontal bar above the dependency relation such as in \texttt{dobj:prefer} and higher-order paths are separated by a dot such as in \texttt{amod.compound:dress}.

Offset representations are the central component in the composition process in the APT framework. Figure 1 shows the APT representations for the adjective \texttt{white} (left) and the APT for the noun \texttt{clothes} (right), as might have been observed in a text collection. Each node holds a multiset of lexemes and the anchor of an APT reflects the current perspective of a lexeme at the given node. An offset representation can be created by shifting the anchor along a given path. For example the lexeme \texttt{white} is at the same node as other adjectives such as \texttt{black} and \texttt{clean}, whereas nouns such as \texttt{shoes or noise} are typically reached via the \texttt{amod} edge.

Offsetting in APTs only involves a change in the anchor, the underlying structure remains unchanged. By offsetting the lexeme \texttt{white} by \texttt{amod} the anchor is shifted along the \texttt{amod} edge, which results in creating a noun view for the adjective \texttt{white}. We denote the offset view of a lexeme for a given path by superscripting the offset path, for example the \texttt{amod} offset of the adjective \texttt{white} is denoted as \texttt{white\textsuperscript{amod}}. The offsetting procedure changes the starting points of the paths as visible in Figure 1 between the anchors for \texttt{white} and \texttt{white\textsuperscript{amod}}, since paths always begin at the anchor. The red dashed line in Figure 1 reflects that anchor shift. The lexeme \texttt{white\textsuperscript{amod}} represents a prototypical “white thing”, that is, a noun that has been modified by the adjective \texttt{white}. We note that all edges in the APT space are bi-directional as exemplified in the coloured \texttt{amod} and \texttt{amod} edges in the APT for \texttt{white}, however for brevity we only show uni-directional edges in Figure 1.

By considering the APT representations for the lexemes \texttt{white} and \texttt{clothes} in Figure 1, it becomes apparent that lexemes with different parts of speech are located in different areas of the semantic space. If we want to compose the adjective-noun phrase \texttt{white clothes}, we need to offset one of the two constituents to align the feature spaces in order to leverage their distributional commonalities. This can be achieved by either creating a noun offset view of \texttt{white}, by shifting the anchor along the \texttt{amod} edge, or by creating an adjective offset representation of \texttt{clothes} by shifting its anchor along \texttt{amod}. In this work we follow Weir et al. (2016) and always offset the dependent in a given relation. Table 1 shows a subset of the features of Figure 1 as would be represented in a vectorised APT. Vectorising the whole APT lexicon results in a very high-dimensional and sparse typed distributional space. The features for \texttt{white\textsuperscript{amod}} (middle column) highlight the change in feature space caused by offsetting the adjective \texttt{white}. The features of the offset view \texttt{white\textsuperscript{amod}}, are now aligned with the noun \texttt{clothes} such that the two can be composed. Composition can be performed by either selecting the \texttt{union} or \texttt{intersection} of the aligned features.

<table>
<thead>
<tr>
<th>\texttt{white}</th>
<th>\texttt{white\textsuperscript{amod}}</th>
<th>\texttt{clothes}</th>
</tr>
</thead>
<tbody>
<tr>
<td>\texttt{:clean}</td>
<td>\texttt{amod:clean}</td>
<td>\texttt{amod:wear}</td>
</tr>
<tr>
<td>\texttt{amod:shoes}</td>
<td>\texttt{:shoes}</td>
<td>\texttt{:dress}</td>
</tr>
<tr>
<td>\texttt{amod.dobj:wear}</td>
<td>\texttt{dobj:wear}</td>
<td>\texttt{dobj:wear}</td>
</tr>
<tr>
<td>\texttt{amod.nsubj:earn}</td>
<td>\texttt{nsubj:earn}</td>
<td>\texttt{nsubj:admit}</td>
</tr>
</tbody>
</table>

Table 1: Sample of vectorised features for the APTs shown in Figure 1. Offsetting \texttt{white} by \texttt{amod} creates an offset view, \texttt{white\textsuperscript{amod}}, representing a noun, and has the consequence of aligning the feature space with \texttt{clothes}.

### 2.1 Qualitative Analysis of Offset Representations

Any offset view of a lexeme is behaviourally identical to a “normal” lexeme. It has an associated part of speech, a distributional representation which locates it in semantic space, and we can find neighbours for it in the same way that we find neighbours for any other lexeme. In this way, a single APT data structure is able to provide many different views of any given lexeme. These views reflect the different ways in which the lexeme is used. For example \texttt{law.nsubj} is the \texttt{nsubj} offset representation of the noun \texttt{law}. This lexeme is a verb and represents an action carried out by the \texttt{law}. This contrasts with \texttt{law.dobj}, which is the \texttt{dobj} offset representation of the noun \texttt{law}. It is also a verb, however represents actions done to the \texttt{law}. Table 2 lists the 10 nearest neighbours for a number of lexemes, offset by \texttt{amod}, \texttt{dobj} and \texttt{nsubj} respectively.

For example, the neighbourhood of the lexeme \texttt{ancient} in Table 2 shows that the offset view for \texttt{ancient\textsuperscript{amod}} is a prototypical representation of an “ancient thing”, with neighbours easily associated with the property \texttt{ancient}. Furthermore, Table 2
Figure 1: Structured distributional APT space. Different colours reflect different parts of speech. Boxes denote the current anchor of the APT, circles represent nodes in the APT space, holding lexemes, and edges represent their relationship within the space.

illustrates that nearest neighbours of offset views are often other offset representations. This means that for example actions carried out by a mother tend to be similar to actions carried out by a father or a parent.

2.2 Offset Inference

Our approach generalises the unsupervised algorithm proposed by Kober et al. (2016), henceforth “standard DI”, as a method for inferring missing knowledge into an APT representation. Rather than simply inferring potentially plausible, but unobserved co-occurrences from near distributional neighbours, inferences can be made involving offset APTs. For example, the adjective white can be offset so that it represents a noun — a prototypical “white thing”. This allows inferring plausible co-occurrences from other “things that can be white”, such as shoes or shirts. Our algorithm therefore reflects the contextualised use of a word. This has the advantage of being able to make flexible and fine grained distinctions in the inference process. For example if the noun law is used as a subject, our algorithm allows inferring plausible co-occurrences from “other actions carried out by the law”. This contrasts the use of law as an object, where offset inference is able to find co-occurrences on the basis of “other actions done to the law”. This is a crucial advantage over the method of Kober et al. (2016) which only supports inference on uncontextualised lexemes.

A sketch of how offset inference for a lexeme w works is shown in Algorithm 1. Our algorithm requires a distributional model M, an APT representation for the lexeme w for which to perform offset inference, a dependency path p, describing the offset for w, and the number of neighbours k. The offset representation of w′ is then enriched with the information from its distributional neighbours by some merge function. We note that if the offset path p is the empty path, we would recover the algorithm presented by Kober et al. (2016). Our algorithm is unsupervised, and agnostic to the input distributional model and the neighbour retrieval function.

Algorithm 1 Offset Inference

1: procedure OFFSET_INFERENCE(M, w, p, k)
2:   w′ ← offset(w, p)
3:   for all n in neighbours(M, w′, k) do
4:     w′′ ← merge(w′′, n)
5:   end for
6: return w′′
7: end procedure

Connection to Distributional Composition

An interesting observation is the similarity between distributional inference and distributional composition, as both operations are realised by the same mechanism — an offset followed by inferring plausible co-occurrence counts for a single lexeme in the case of distributional inference, or for a phrase in the case of composition. The merging of co-occurrence dimensions for distributional inference can also be any of the operations commonly used for distributional composition such as pointwise minimum, maximum, addition or multiplication.

This relation creates an interesting dynamic between distributional inference and composition when used in a complementary manner as in this work. The former can be used as a process of co-occurrence embellishment which is adding missing information, however with the risk of introducing some noise. The latter on the other hand can be
Table 2: List of the 10 nearest neighbours of amod, dobj and nsubj offset representations.

<table>
<thead>
<tr>
<th>Offset Representation</th>
<th>Nearest Neighbours</th>
</tr>
</thead>
<tbody>
<tr>
<td>ancient</td>
<td>civilization, mythology, tradition, ruin, monument, trackway, tomb, antiquity, folklore, deity</td>
</tr>
<tr>
<td>red</td>
<td>blue, black, green, dark, onion, pepper, red, tomato, carrot, garlic</td>
</tr>
<tr>
<td>economic</td>
<td>political, societal, cohabiting, economy, growth, cohabitant, globalisation, competitiveness, globalization, prosperity</td>
</tr>
<tr>
<td>government</td>
<td>overthrow, party, authority, leader, capital, force, state, official, minister, oust</td>
</tr>
<tr>
<td>law</td>
<td>difficulty, solve, coded, issue, injury, overcome, question, think, loss, relieve</td>
</tr>
<tr>
<td>researcher</td>
<td>physician, writer, theorize, thwart, theorise, hypothesize, surmise, student, worker, apprehend</td>
</tr>
<tr>
<td>mother</td>
<td>wife, father, parent, woman, re-married, remarry, girl, breastfeed, family, disown</td>
</tr>
<tr>
<td>law</td>
<td>rule, principle, policy, criminalize, case, contract, prohibit, proscribe, enjoin, charge</td>
</tr>
</tbody>
</table>

used as a process of co-occurrences filtering, that is leveraging the enriched representations, while also sieving out the previously introduced noise.

3 Experiments

For our experiments we re-implemented the standard DI method of Kober et al. (2016) for a direct comparison. We built an order 2 APT space on the basis of the concatenation of ukWaC, Wackypedia and the BNC (Baroni et al., 2009), pre-parsed with the Malt parser (Nivre et al., 2006). We PPMI transformed the raw co-occurrence counts prior to composition, using a negative SPPMI shift of $\log 5$ (Levy and Goldberg, 2014b). We also experimented with composing normalised counts and applying the PPMI transformation after composition as done by Weeds et al. (2017), however found composing PPMI scores to work better for this task.

We evaluate our offset inference algorithm on two popular short phrase composition benchmarks by Mitchell and Lapata (2008) and Mitchell and Lapata (2010), henceforth ML08 and ML10 respectively. The ML08 dataset consists of 120 distinct verb-object (VO) pairs and the ML10 dataset contains 108 adjective-noun (AN), 108 noun-noun (NN) and 108 verb-object pairs. The goal is to compare a model’s similarity estimates to human provided judgements. For both tasks, each phrase pair has been rated by multiple human annotators on a scale between 1 and 7, where 7 indicates maximum similarity. Comparison with human judgements is achieved by calculating Spearman’s $\rho$ between the model’s similarity estimates and the scores of each human annotator individually. We performed composition by intersection and tuned the number of neighbours by a grid search over $\{0, 10, 30, 50, 100, 500, 1000\}$ on the ML10 development set, selecting 10 neighbours for NNs, 100 for ANs and 50 for VOs for both DI algorithms. We calculate statistical significance using the method of Steiger (1980).

**Effect of the number of neighbours**

Figure 2 shows the effect of the number of neighbours for AN, NN and VO phrases, using offset inference, on the ML10 development set. Interestingly, NN compounds exhibit an early saturation effect, while VOs and ANs require more neighbours for optimal performance. One explanation for the observed behaviour is that up to some threshold, the neighbours being added contribute actually missing co-occurrence events, whereas past that threshold distributional inference degrades to just generic smoothing that is simply compensating for sparsity, but overwhelming the representations with non-plausible co-occurrence information. A similar effect has also been observed by Erk and Pado (2010) in an exemplar-based model.

![Figure 2: Effect of the number of neighbours on the ML10 development set.](image)

**Results**

Table 3 shows that both forms of distributional inference significantly outperform a baseline without DI. On average, offset inference outperforms the method of Kober et al. (2016) by a statistically...
significant margin on both datasets.

<table>
<thead>
<tr>
<th>APF configuration</th>
<th>ML10</th>
<th>ML08</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AN</td>
<td>SN</td>
</tr>
<tr>
<td>None</td>
<td>0.35</td>
<td>0.30</td>
</tr>
<tr>
<td>Standard DI</td>
<td>0.48</td>
<td>0.51</td>
</tr>
<tr>
<td>Offset Inference</td>
<td>0.49</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Table 3: Comparison of DI algorithms. ‡ denotes statistical significance at $p < 0.01$ in comparison to the method without DI, * denotes statistical significance at $p < 0.01$ in comparison to standard DI and † denotes statistical significance at $p < 0.05$ in comparison to standard DI.

Table 4 shows that offset inference substantially outperforms comparable sparse models by Dinu et al. (2013) on ML08, achieving a new state-of-the-art, and matches the performance of the state-of-the-art neural network model of Hashimoto et al. (2014) on ML10, while being fully interpretable.

Model | ML10 - Average | ML08 |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Our work</td>
<td>0.48</td>
<td>0.31</td>
</tr>
<tr>
<td>Blacoe and Lapata (2012)</td>
<td>0.44</td>
<td>-</td>
</tr>
<tr>
<td>Hashimoto et al. (2014)</td>
<td>0.48</td>
<td>-</td>
</tr>
<tr>
<td>Weir et al. (2016)</td>
<td>0.43</td>
<td>0.26</td>
</tr>
<tr>
<td>Dinu et al. (2013)</td>
<td>-</td>
<td>0.23 – 0.26</td>
</tr>
<tr>
<td>Erk and Padó (2008)</td>
<td>-</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Table 4: Comparison with existing methods.

4 Related Work


Composition with distributional semantic models has become a popular research area in recent years. Simple, yet competitive methods, are based on pointwise vector addition or multiplication (Mitchell and Lapata, 2008, 2010). However, these approaches neglect the structure of the text defining composition as a commutative operation.

A number of approaches proposed in the literature attempt to overcome this shortcoming by introducing weighted additive variants (Guevara, 2010, 2011; Zanzotto et al., 2010). Another popular strand of work models semantic composition on the basis of ideas arising in formal semantics. Composition in such models is usually implemented as operations on higher-order tensors (Baroni and Zamparelli, 2010; Baroni et al., 2014; Coo-}

5 Conclusion

In this paper we have introduced a novel form of distributional inference that generalises the method introduced by Kober et al. (2016). We have shown its effectiveness for semantic composition on two benchmark phrase similarity tasks where we achieved state-of-the-art performance while retaining the interpretability of our model. We have furthermore highlighted an interesting connection between distributional inference and distributional composition.

In future work we aim to apply our novel method to improve modelling selectional preferences, lexical inference, and scale up to longer phrases and full sentences.
References


Richard Socher, Andrej Karpathy, Quoc Le, Christopher Manning, and Andrew Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association for Computational Linguistics 2:207–218.


