Words, Concepts, and the Geometry of Analogy

Stephen McGregor, Matthew Purver, and Geraint Wiggins
Queen Mary University of London
SLPCS, June 11 2016
University of Strathclyde, Glasgow

The Contextual Modelling of Concepts

This presentation will outline a proposal for a way towards using a word-counting approach to computational linguistics for building dynamic, context-sensitive geometric models of conceptual relationships, focusing on analogy. I'll be covering:

- An overview of an ostensibly geometric method for discovering anlogies in semantic spaces, and a survey of some problems
- A novel distributional semantic model for the extemporaneous projection of conceptually contextualised subspaces
- A proposal for a new methodology for geometrically mapping analogical relationships in context specific subspaces

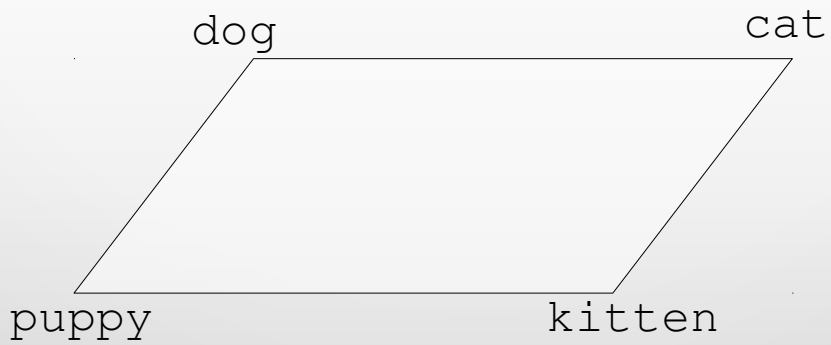
This is very much a work-in-progress, exploring ways to extend early results in this context sensitive approach to lexical modelling towards discovering more robustly geometric conceptual spaces.

The Geometry of Analogy

Beginning with the problem of modelling analogy geometrically in semantic spaces, the approach proposed by Mikolov et al (2013) suggests that, in a distributional semantic space, analogies can be expressed algebraically:

$$\overrightarrow{\text{cat}} - \overrightarrow{\text{dog}} + \overrightarrow{\text{puppy}} = \overrightarrow{\text{kitten}}$$

The implication is that these analogical relationships can be discovered as parallelograms sitting on oblique planes within the high dimensional space of the model.



Hits and Misses

This approach has proven productive...

- *France* is to *Paris* as *Finland* is to ***Helsinki*** ✓
- *girl* is to *boy* as *prince* is to ***princess*** ✓
- *fast* is to *faster* as *big* is to ***bigger*** ✓

(*Results here produced using Mikolov et al's vectors, trained on a Google News corpus containing about 6 billion word tokens using their word2vec model.)

Hits and Misses

This approach has proven productive...

- *France* is to *Paris* as *Finland* is to ***Helsinki*** ✓
- *girl* is to *boy* as *prince* is to ***princess*** ✓
- *fast* is to *faster* as *big* is to ***bigger*** ✓

...to a point.

- *elephant* is to *mouse* as *big* is to ***huge*** ✗
- *bank* is to *river* as *shoulder* is to ***elbow*** ✗
- *picture* is to *paint* as *story* is to ***mexican viagra viagra*** ✗

Hits and Misses

This approach has proven productive...

- *France* is to *Paris* as *Finland* is to ***Helsinki*** ✓
- *girl* is to *boy* as *prince* is to ***princess*** ✓
- *fast* is to *faster* as *big* is to ***bigger*** ✓

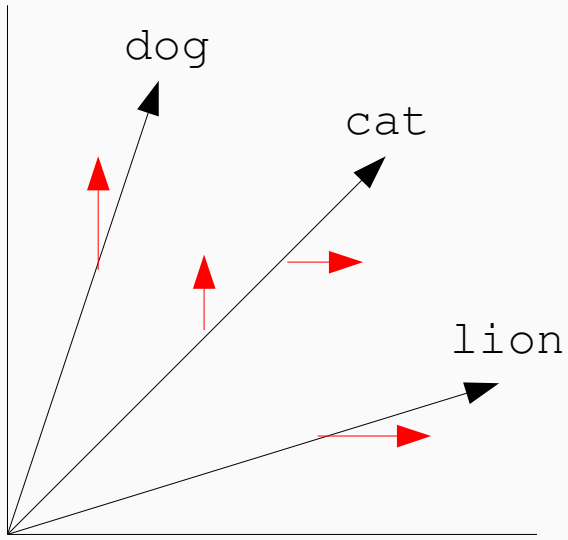
...to a point.

- *elephant* is to *mouse* as *big* is to ***huge*** ✗
- *bank* is to *river* as *shoulder* is to ***elbow*** ✗
- *picture* is to *paint* as *story* is to ***mexican viagra viagra*** ✗

It's not hard to predict the cases where these analogies fail to emerge in a general space of semantic relationships. By picking conceptual relationships that are characterised by relativism, referential ambiguity, metaphor, and the like – in a word, by context – we can trip the model up.

The Utility of Literal Dimensions

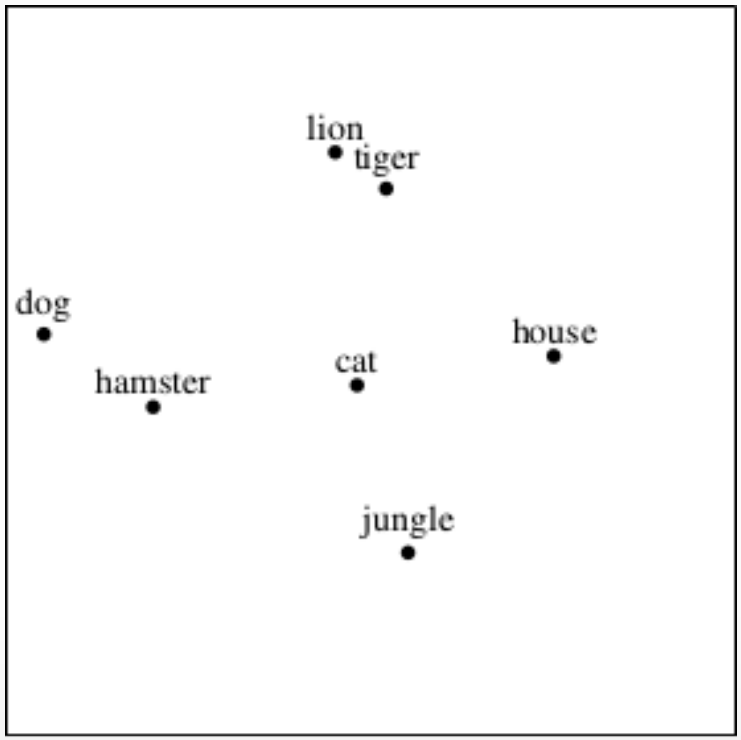
In a space trained by a neural network or derived from matrix factorisation, dimensions are, by design, maximally informative, and there is therefore no recourse to a context specific projection of a subspace. In these spaces, dimensions exist just for the sake of moving vectors.



We propose that in a space of dimensions bearing literal co-occurrence values, there should be recourse to some lower-dimensional projection where analogies can be discovered as geometric regions within a subspace.

Spaces of Meaning

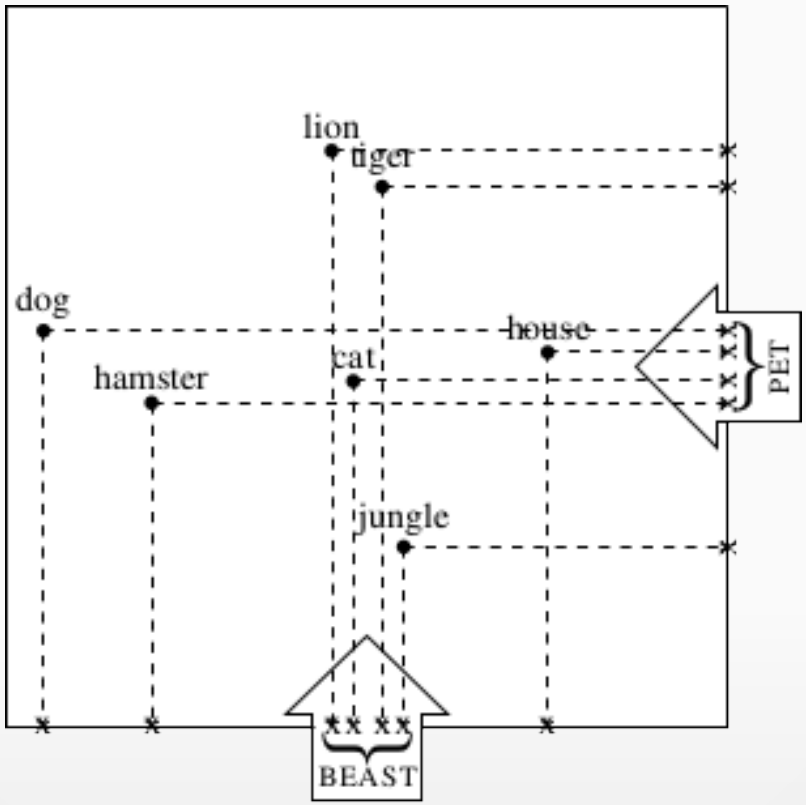
We begin by building a space of word-vectors based on a typical PMI type metric of word co-occurrences, based on a traversal of English language Wikipedia.



As expected, this technique generates a space of word vectors where semantic relatedness is indicated by proximity—but, also predictably, this space is messy.

Contextualised Spaces of Meaning

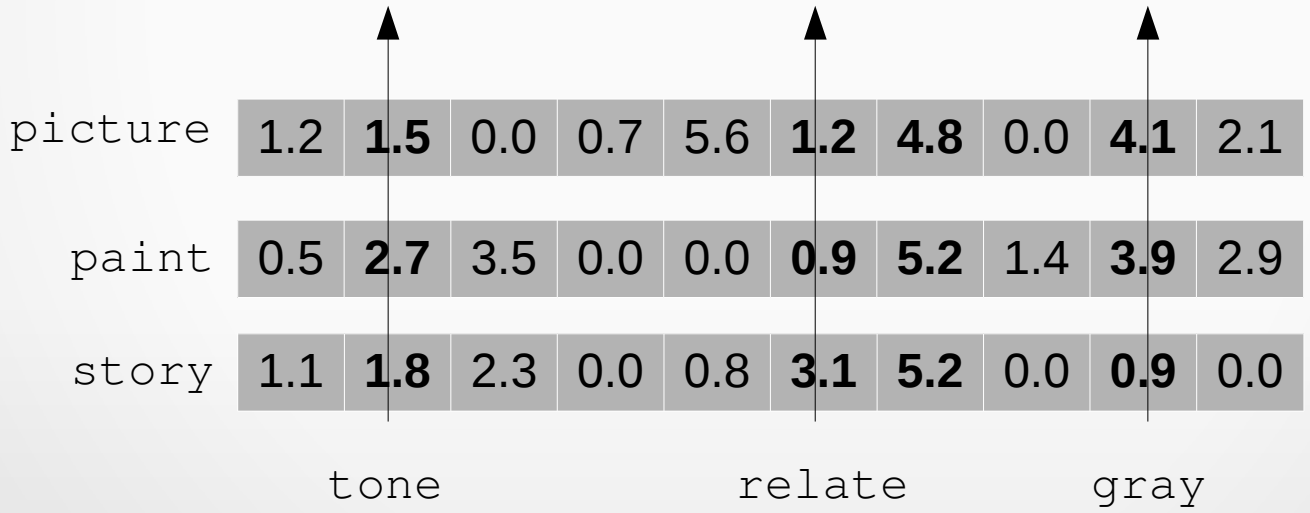
By projecting lower-dimensional subspaces on the base semantic space, however, we can chose perspectives, so to speak, where words cluster in contextually meaningful ways.



Finding Analogical Contexts

Returning to the problem of analogy, we can now try to find subspaces of our very sparse base space where analogical relationships are mapped geometrically.

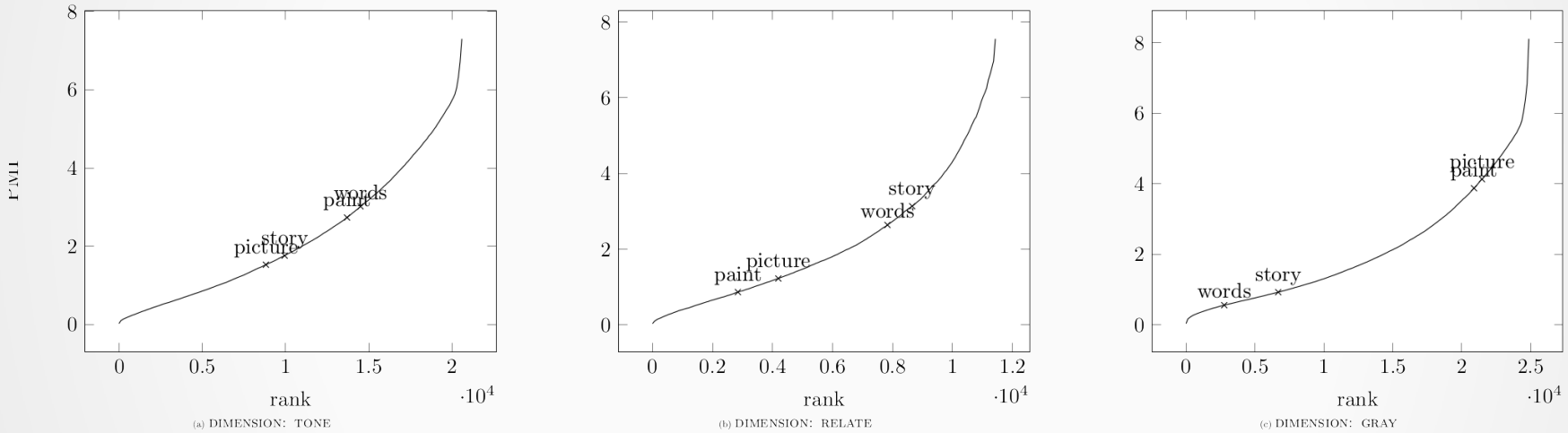
In order to test the model, we've held back 1,000 analogies from the test set provided by Google. For each analogy A:B::C:D, we build a 200 dimensional subspace consisting of the collectively non-zero dimensions with the highest mean PMI value for A, B, and C.



From these 200 dimensions, we then select the 20 dimensions on which the relationship A-B = C-D is most nearly fulfilled.

Word Clusters in Context

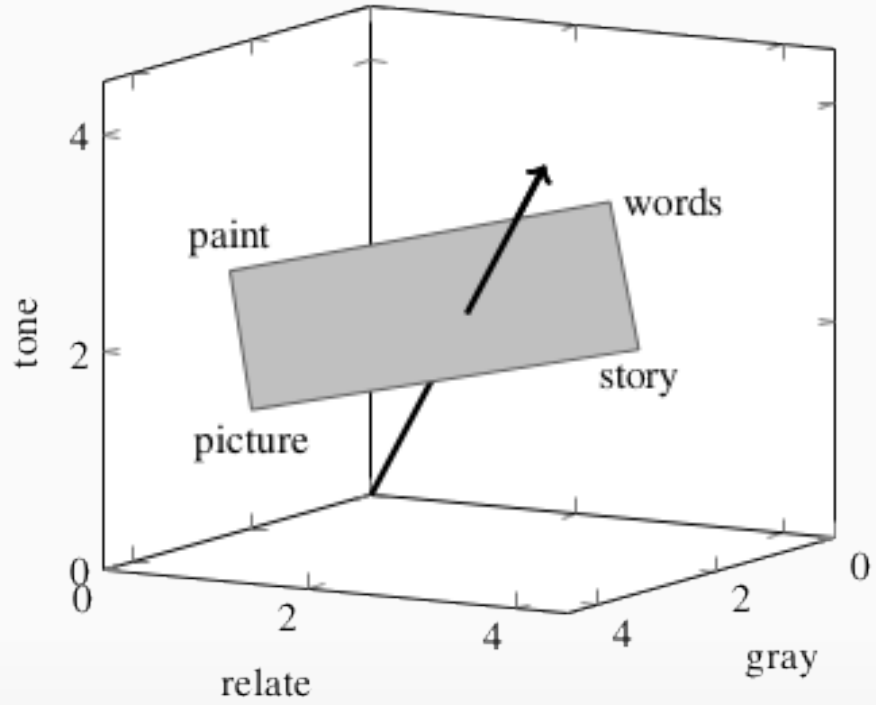
Of the 1,000 held back analogies, 3 contained words that were outside out model's vocabulary. Of the remaining 997, 996 satisfied the equation B-A+C ≈ D.



These subspaces can be understood as contexts in which the analogy emerges, corresponding to some environmental situation. And we can observe clustering tendencies that play out across the analogically productive dimensions delineating such a subspace.

The Emergent Geometry of Analogy

Given these observations, we can predict some salient aspects of the geometry of the regions that will delineate analogies.



In particular, we expect these regions to form parallelograms that sit centrally in the positive region of the contextual subspace, situated more or less perpendicularly to a line extending from the origin through the centre of the space.

References and Acknowledgement

Joaquín Derrac & Steven Schockaert (2014): *Characterising Semantic Relatedness Using Interpretable Diirections in Conceptual Spaces.* In: 2nd European Conference on Artificial Intelligence, pp. 243–248.

Tomas Mikolov, Kai Chen, Greg Corrado & Jeffrey Dean (2013): *Efficient Estimation of Word Representations in Vector Space.* In: Proceedings of ICLR Workshop.

Laura Rimell (2014): *Distributional Lexical Entailment by Topic Coherence.* In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg.

EPSRC
Engineering and Physical Sciences Research Council

This research has been supported by EPSRC grant EP/L50483X/1.