What does it all mean?

Compositional Distributional Semantics for Modelling Natural Language

Thomas Kober
t.kober@sussex.ac.uk

PyData Berlin

2nd July 2017 (1498996800)

Whats this all about?

- Well…thats a good question, glad you asked!
- If you’ve ever added word embeddings (e.g. from word2vec) together, or used them as input to a neural net…
- …you’ve applied a composition function to distributional word representations
- This talk is intended to give you some background on the current state of the research in that area
    - Overview of why it is useful
    - Emphasis on its current limitations
- Its 💀dangerously academic💀 at times
- But shouldn’t be too bad (I hope!)

Outline

- Compositional Distributional Semantics 101
- Distributional word representations (and why they are cool)
- Composition - A small overview
- Composition - Its complicated…
- Applications

Compositional Distributional **Semantics**

- **Semantics** - The study of the meaning of words and phrases in a language

Compositional
Distributional Semantics

- **Distributional** - Based on the co-occurrence statistics of words in a corpus

- **Semantics** - The study of the meaning of words and phrases in a language

Compositional Distributional Semantics

- **Compositional** - Based on the product of combining elementary word representations

- **Distributional** - Based on the co-occurrence statistics of words in a corpus

- **Semantics** - The study of the meaning of words and phrases in a language

Word Representations

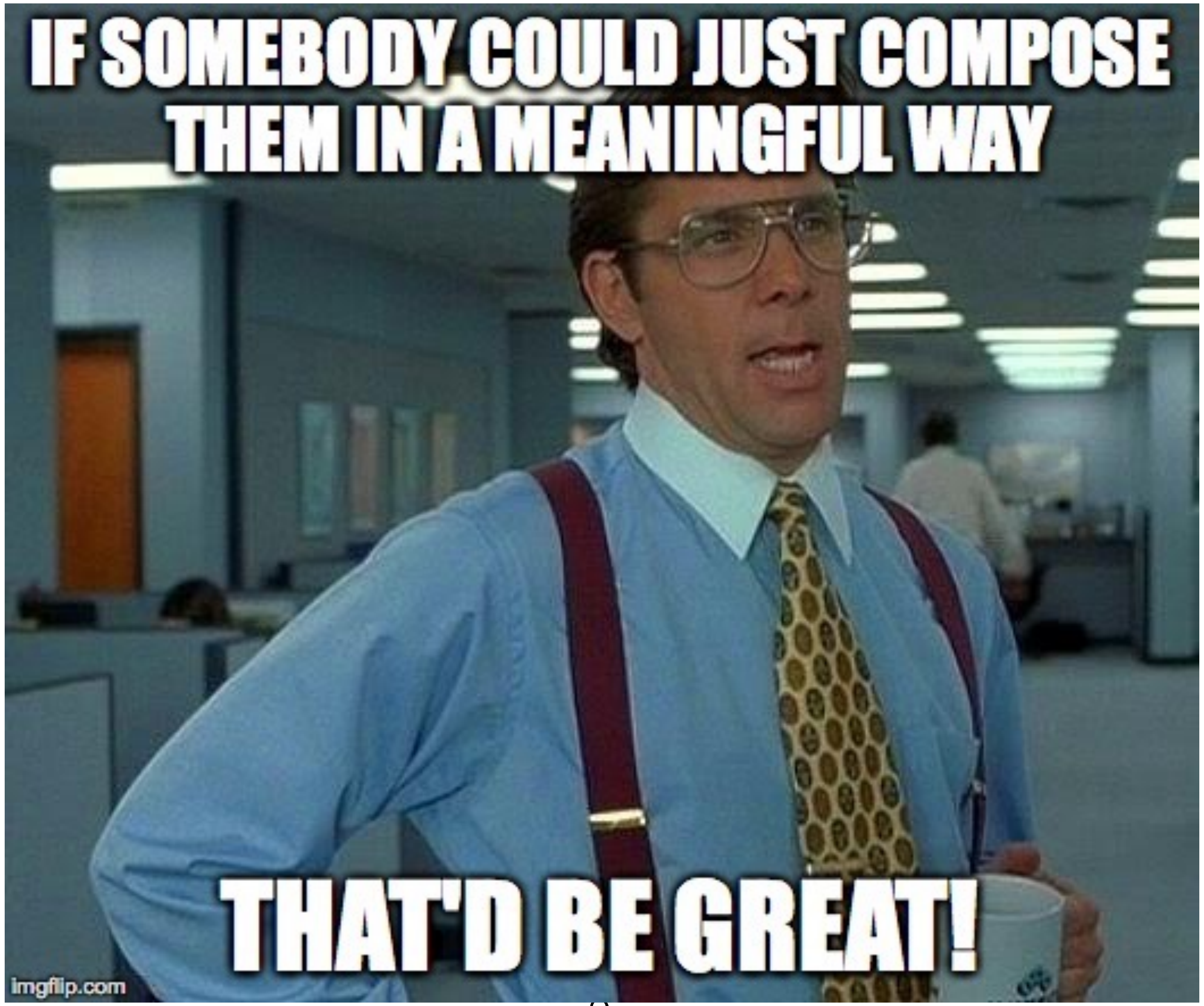
- Distributional Hypothesis
  - Similar words tend to occur in similar contexts - Harris (1954)
  - "*You shall know the meaning of a word by the company it keeps*" - Firth (1962)
- Long history in NLP research
  - e.g. Sparck-Jones (1986), Church and Hanks (1989), Deerwester et al. (1990)
  - Continuous model of word meaning
  - Words are represented in a high-dimensional metric space
- Count vs. predict (Baroni et al., 2014)
  - Explicitly counting co-occurrences, e.g. PPMI based word representations or GloVe
  - Context predicting models, e.g. word2vec
- All models based on the underlying co-occurrence statistics in a corpus

Why we ♡ them

- Capture interesting linguistic regularities
    - **v**(king) **-** **v**(man) **+** **v**(woman) **≈** **v**(queen)
- Can measure semantic similarity between words (more powerful than it sounds)
- Unsupervised algorithm scalable to large corpora with billions of tokens
    - Also language agnostic!
- Plug and Play
    - Download pre-trained ones, roll your own with **gensim**, etc.
    - Add to your NLP pipeline, sit back and relax
- Flexible
    - Can use the off-the-shelf ones as drop-in - No need to train them with a task
    - But you can if you need to

♫ Composing words ♫

- Lots of effort into creating word representations, but…
- …they are just single words!

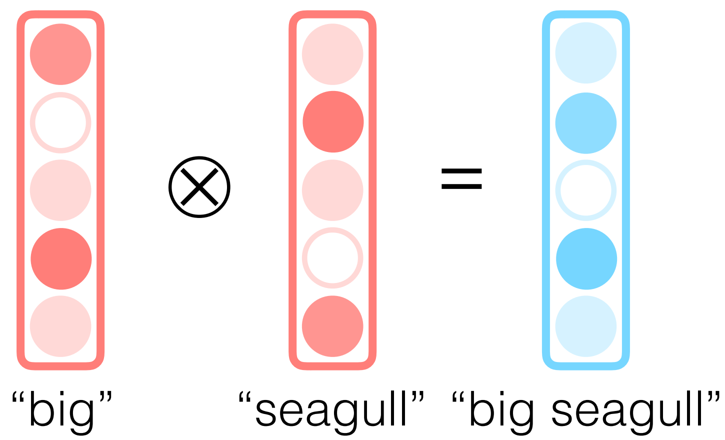


♫ Composing words ♫

- Lots of effort into creating word representations, but…
- …they are just single words!
- Would be nice if there was an off-the-shelf component that creates meaningful representation of longer phrases and sentences
- Some plug-and-play composition function that integrates effortlessly with distributional word representations
- Same level of flexibility
    - Option to use as-is or fine-tune for a given task
- 4 major approaches to modelling distributional composition
    - Pointwise addition/multiplication
    - Semantic composition based on Formal Semantics
    - Anchored Packed Trees
    - Neural Networks

We’ve got this far by now

- Pointwise addition/multiplication



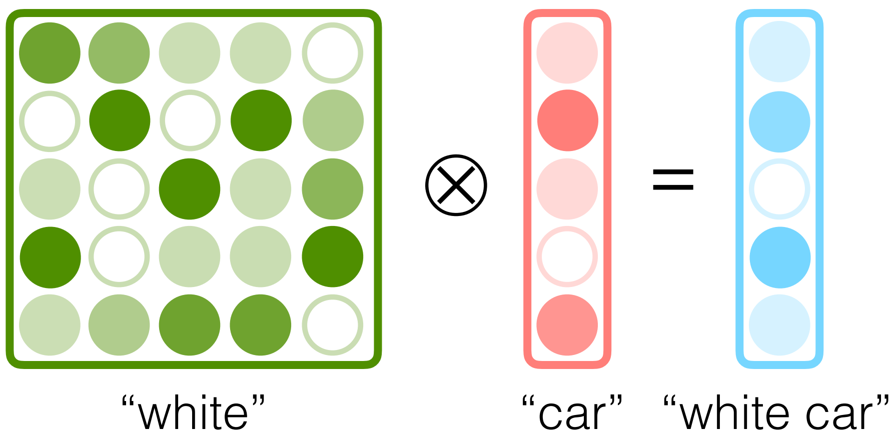
Major Problem - commutativity:

$\mathbf{v}(\mathit{race}) + \mathbf{v}(\mathit{car}) = \mathbf{v}(\mathit{car}) + \mathbf{v}(\mathit{race})$

- Often represents an annoyingly-hard-to-beat baseline (Blacoe and Lapata, 2012; Hill et al., 2016)
- Despite their simplicity capture some interesting patterns
    - Pointwise multiplication in explicit PPMI vectors represents a (weighted) feature intersection
    - So does pointwise addition in neural word embeddings (Tian et al., 2017)
    - Achieves contextualisation and is able to recover sense specific information remarkably well (Kober et al., 2017)
- Commutativity not so problematic for some tasks (e.g. Text Classification) as for others (e.g. Recognising Textual Entailment)

We've got this far by now

- Formal Semantics
    - Based on the notion of function application
    - e.g. an adjective is a function acting on a noun



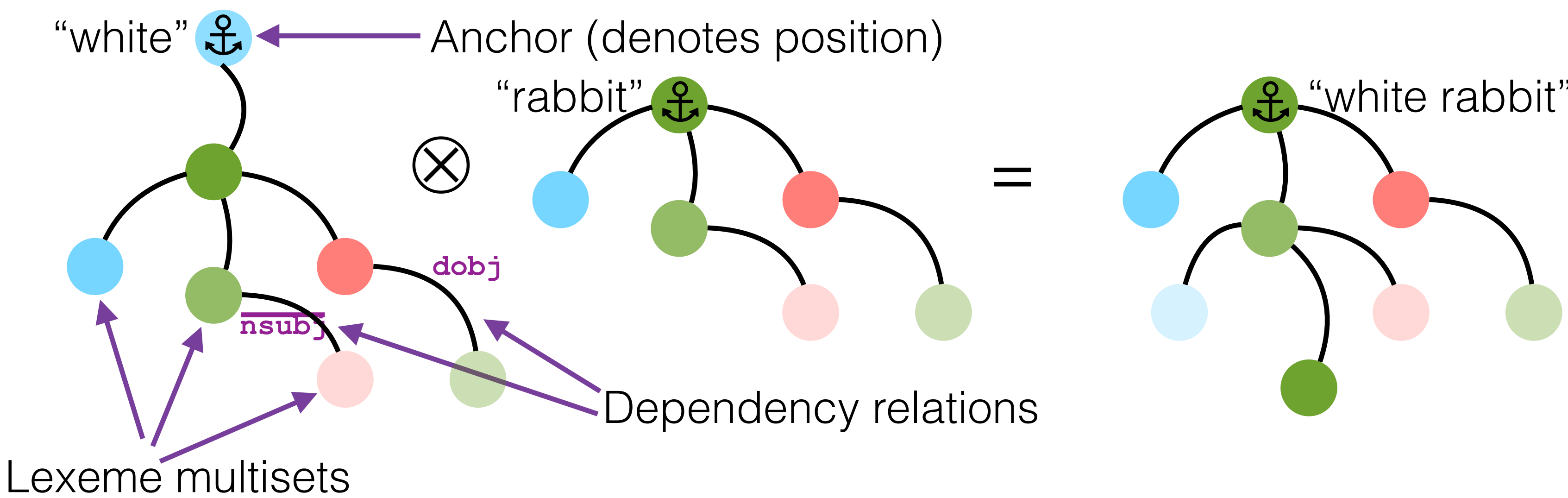
Major Problem - scalability:

Order of a word representation depends on its category
e.g. a verb would be a 3rd order Tensor

- Theoretically sound and linguistically grounded approach (Baroni and Zamparelli, 2010; Coecke et al., 2011)
- Sentences of different lengths often end up with different dimensionality
    - How to calculate similarity between them?
- Very difficult to scale beyond short sentences

We’ve got this far by now

- Anchored Packed Trees (Weir et al., 2016)



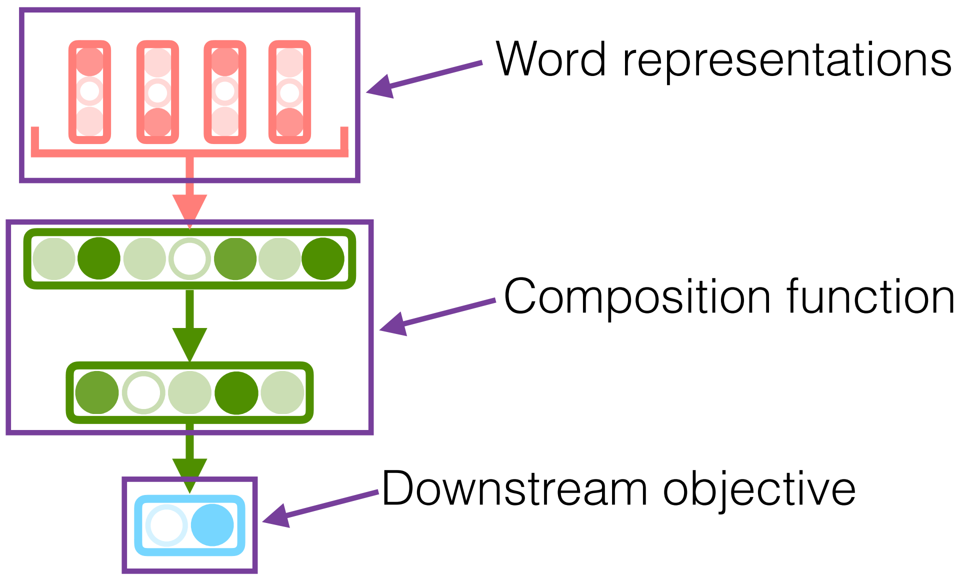
Major Problem - sparsity:

co-occurrences are typed
dimensionality of the space explodes

- Based on a dependency parsed corpus
- Nodes are weighted lexeme multisets
- Edges are dependency relations as observed in the corpus
- Composition involves an additional step - offsetting - to align incompatible representations

We've got this far by now

- Neural Networks



Major Problem - transferability:

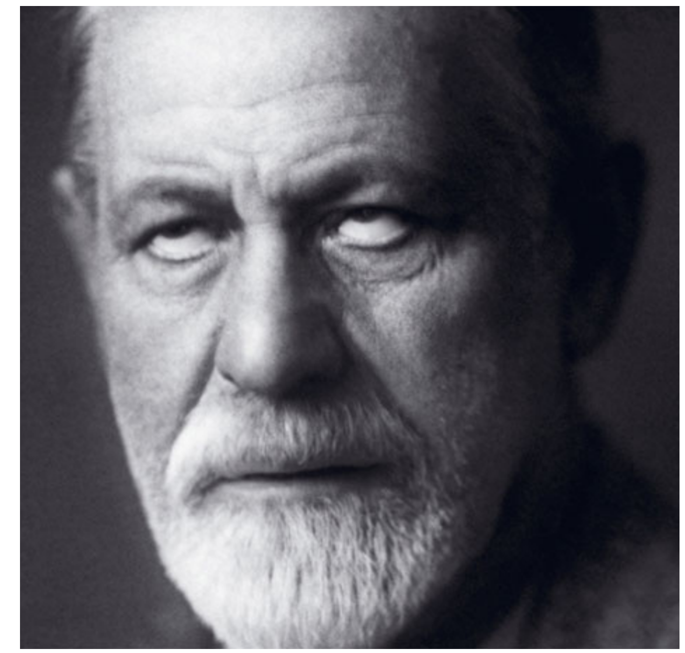
Composition function is trained for a specific task
Plugging learnt model into another task often fails
(Mou et al., 2016)

- Composition function learnt as part of an end-to-end model from data and…
- …tailored to a given task
- Different tasks require different composition functions
- Not general purpose or plug and play because need to be re-trained with the given task
  - But currently lots of progress in Multi-Task Learning and Domain Adaption
- Despite the tailoring and computational effort often achieve only small improvements over just adding word representations (e.g. Iyyer et al., 2015; Wieting et al., 2016)

We’ve got this far by now

- To summarise…
- Its not all that bad
- All major approaches have “issues”
    - Composition functions are not yet such a nice & general purpose drop-in as word representations
- But they are reasonably practical and useful, however there’s more problems…
- …which one is the best and how to measure this?
- One obstacle that potentially slows down progress is a good way to evaluate and compare composition functions

More Problems: Evaluation

- Evaluation either based on a phrase similarity task in comparison to human judgements
    - Similarity is already a difficult on the lexical level - it doesn't get easier with more words…
- Or based on the performance of a downstream task
    - Too many factors that can influence performance
- Difficult to design a "good" task, need to figure out what we actually want to achieve
    - Generality of a composition function across downstream tasks (and without re-training)?
    - Paraphrasing? Entailment?
    - Is it actually task specific?

What does it *actually* all mean?

- Even if everything would be working perfectly, there are some broader issues
- What does a sentence actually mean?
    - The longer the sentence the more difficult it becomes
- Whats the meaning of the following sentence?
    - "*The battle ended at nightfall, with the victory remaining a matter of opinion: that the Parliamentarian foot were still in position at nightfall when, as the Royalists themselves admitted, they drew back a little; or that next morning the Royalists occupied the field after the Parliamentarians retreated in the night.*"

What does it *actually* all mean?

- Should a sentence like this really be encoded in a single vector?
    - Ray Mooney: "*You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector*"
- The problem is not just philosophical but also practical (Polajnar et al., 2014)
    - Intersective composition functions (e.g. pointwise multiplication) run out of overlapping features at some point - the result is an empty vector
    - Composition by union accumulates too much information - can't discriminate anything from anything

There are some good sides to it

- Leverage existing resources effectively
    - Word representations are great
- Contextualisation
    - Word representations usually a weighted sum of all the different usages of word (Arora et al., 2016)
    - Composition has a sense discriminating effect
    - Given some context, polysemy might not be such a problem
    - Even simple composition function can recover a non-trivial amount of sense specific information (Kober et al., 2017)
- Successful component in many different systems (Parsing, MT, Sentiment Analysis, QA, …)
- Plug & Play and General Purpose?
    - Yeah, maybe tomorrow…
- Briefly look at two applications
    - Aspect based Sentiment Analysis
    - Question Answering

Aspect based Sentiment Analysis

- Not just interested in overall sentiment of a review, but in specific aspects
    - For a camera, say the lens or the battery or the weight or whatever
- For analysing sentiment of a full review, bag-of-words + TF-IDF + SVM is probably good enough
    - For specific aspects, we need a more fine grained understanding on the sentence level
    - Can give a more detailed insight of what makes a review 3/5 or 4/5 instead of 5/5
    - Interesting problem - product is being liked, but there was something that was unsatisfactory
- Composition represents a central part of a larger system
    - Create compositional representation of sentences (e.g. Alghunaim et al., 2015)
    - Compositional sentence representation = continuous scale of similarity
    - Can create multiple representations of a sentence to allow inferences w.r.t. multiple aspects
    - Good way to identify issues that are being talked about a lot
    - Understanding what individuals look for in a product can also help to improve product recommendation

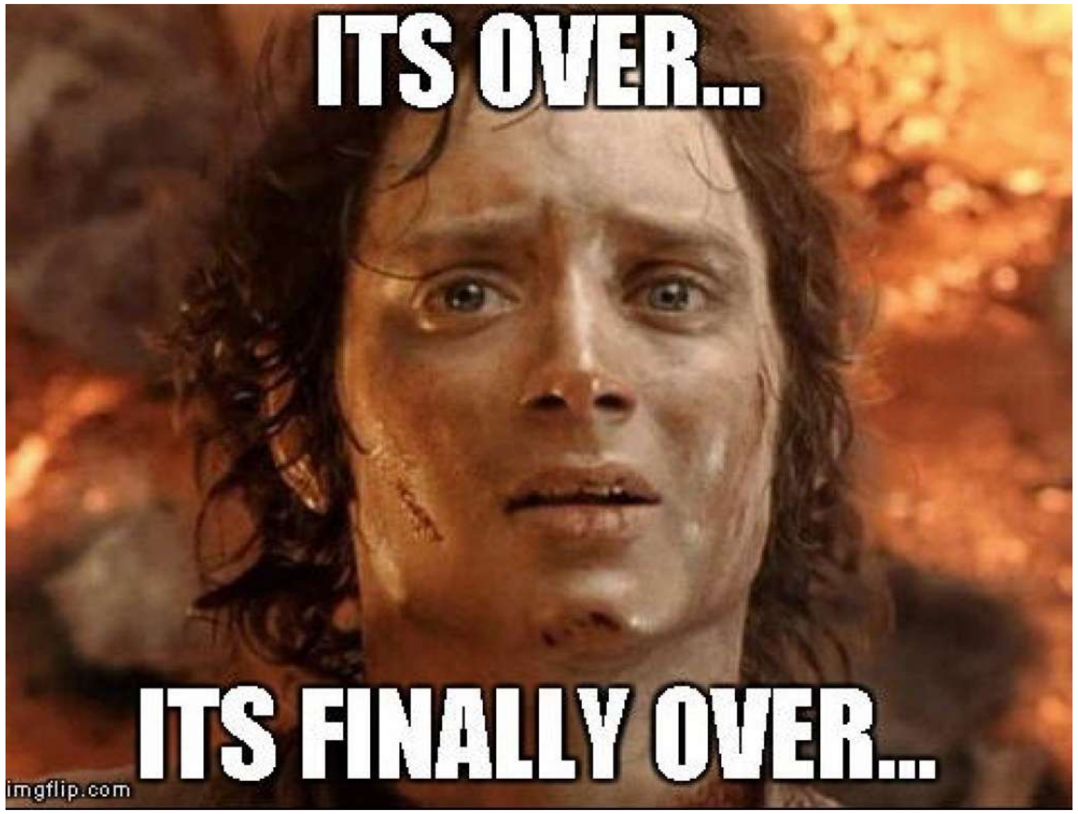
Question Answering

- Traditionally, setup as an information retrieval problem
    - Goal is to retrieve the most relevant answers to a given question
- Task setup usually has ***n*** answer candidates for a given question (e.g. Iyyer et al., 2014)
- Exploit distributional representations of answers and questions by leveraging their commonalities
    - Sharing of semantic content
    - Composition achieves contextualisation of content words, and acts as a mechanism to effectively integrate distributional knowledge into a representation
    - "*What was the name of the fascist dictator of Italy during WWII?*"
        - a) *Walt Disney*?
        - b) *Rhianna*?
        - c) *Benito Mussolini*?
    - Expect to be more semantic overlap of the composed representation of the question with the correct answer c) than with the incorrect ones
- Not restricted to simple named entity style questions

Summary

- Distributional word representations are great and composition is a way to effectively leverage these existing resources
- Composing word representations does work but has its limitations
- Different composition functions have different shortcomings
    - unsupervised & general vs. supervised & specific
    - How to evaluate them?
- Lots of research going on and lots of progress being made

Thats it!



email: t.kober@sussex.ac.uk

strong opinions in 140+ chars: **@tttthomasssss**

buggy code: **github.com/tttthomasssss**

References (1)

- Abdulaziz Alghunaim, Mitra Mohtarami, Scott Cyphers and Jim Glass. 2015. A Vector Space Approach for Aspect Based Sentiment Analysis. In Proceedings of the 1st Workshop on Vector Space Modelling for Natural Language Processing, 116-122

- Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma and Andrej Risteski. 2016. Linear Algebraic Structure of Word Senses, with Applications to Polysemy, arXiv:1601.03764

- Marco Baroni and Roberto Zamparelli. 2010. Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space. In Proceedings of EMNLP, 1183-1193

- Marco Baroni, Georgina Dinu and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of ACL, 238-247

- William Blacoe and Mirella Lapata. 2012. A Comparison of Vector-based Representations for Semantic Composition. In Proceedings of EMNLP, 546-556

- Kenneth Ward Church and Patrick Hanks. 1989. Word Association, Mutual Information, and Lexicography. In Proceedings of ACL, 76-83

- Bob Coecke, Mehrnoosh Sadrzadeh and Stephen Clark. 2011. Mathematical Foundations for a Compositional Distributed Model of Meaning. Linguistic Analysis, 36(1-4): 345-384

- Scott Deerwester, Susan Dumais, George Furnas, Thomas Landauer and Richard Harshman. 1990. Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6): 391-407

- John Rupert Firth. 1962. A Synopsis of Linguistic Theory. Selected Papers of JR Firth 1952-1959, 168-205

References (2)

- Zellig Harris. 1954. Distributional Structure. Word 10:146-162
- Felix Hill, KyungHyun Cho, Anna Korhonen and Yoshua Bengio. 2016. Learning to Understand Phrases by Embedding the Dictionary. TACL 2016(4): 17-30
- Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher and Hal Daumé III. 2014. A Neural Network for Factoid Question Answering over Paragraphs. In Proceedings of EMNLP, 633-644
- Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber and Hal Daumé III. 2015. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In Proceedings of ACL, 1681-1691
- Thomas Kober, Julie Weeds, John Wilkie, Jeremy Reffin and David Weir. 2017. One Representation per Word - Does it make Sense for Composition? In Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications, 79-90
- Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang and Zhi Jin. 2016. How Transferable are Neural Networks in NLP Applications? In Proceedings of EMNLP, 479-489
- Tamara Polajnar, Laura Rimell and Stephen Clark. 2014. Evaluation of Simple Distributional Compositional Operations on Longer Texts. In Proceedings of LREC, 4440-4443
- Karen Sparck-Jones. 1986. Synonymy and Semantic Classification. Edinburgh University Press
- Ran Tian, Naoaki Okazaki and Kentaro Inui. 2017. The mechanism of additive composition. Machine Learning 106(7): 1083-1130
- David Weir, Julie Weeds, Jeremy Reffin and Thomas Kober. 2016. Aligning Packed Dependency Trees: A theory of composition for distributional semantics. Computational Linguistics 42(4):727-761
- John Wieting, Mohit Bansal, Kevin Gimpel and Karen Livescu. 2016. Towards Universal Paraphrastic Sentence Embeddings. In