Seminar 2: Distributional representations

S. Clark. Vector space models of lexical meaning. In S. Lappin and C. Fox, editors, Handbook of Contemporary Semantics — second edition, chapter 16, pages 493–522. Wiley – Blackwell, 2015.

Questions to discuss

On vector space models:

- What notion of meaning is represented by distributional representations?
  - What semantic relations do they capture?
  - How do these relate to the semantic relations we intuitively recognise in natural language?
  - Are there relations that they do not capture?
  - Think of examples in natural language that can modelled well with distributional relations and examples that cannot be.
- How does this notion of meaning different from that taken in model-theoretic semantics that we looked at earlier?
  - Sense and reference?
- What are the main ... for representing meaning of natural language this way?
  - benefits
  - challenges
  - limitations (and dangers!)
- What computational resources, tools and methods do we use to create these representations?
- For what tasks can we use these representations? For what tasks we cannot use them?
- What would be alternative representations?

On compositionality:

- What are the reasons and benefits of combining formal representations with distributional ones?
- What do you think are the biggest challenges of such hybrid models and representations?
- To what degree can we interpret distributional representations?
  - How does this relate to how well a mapping between two types of representations can be achieved?
- There are several different ways to write a formal grammar. How would this affect the mapping?

Simon's notes

- Lexical semantics, can process information from texts without the need to know the model
- Compare similarity between words

- Origins in linguistics and cognitive science (memory, priming)
- Vector space models and information retrieval
- Basis vector identify dimensions (in comparison to the actual vectors for words that we compare)
- TF/IDF: counter the problem that some words are generally more frequent than others, divide by the number of documents the word occurs in (or multiply by the 1/DF
  - see Karen Spärck Jones
- PMI: another method that measures the strength of association: how often two events co-occur compared to how often they occur independently, see J&M Chapter 6; related to log-likelihood ratio and KL-Divergence, see here https://stats.stackexchange.com/questions/179010/difference-between-pointwise-mutual-information-and-log-likelihood-ratio
- Normalise the dot product by the length of the documents and hence we get an equation for cosine
  - compare with the tensor product later
  - dot product gives us a single value
- Dimensionality reduction: singular value decomposition = latent semantic analysis/indexing, non-negative matrix factorisation, random indexing: identifying meaningful dimensions, reduces noise, helps with sparsity
- syntagmatic and paradigmatic relations between words, relevance for contexts
- Curran investigates different ways to represent contexts, including POS tags and relations, linguistic processing words such as stemming, removal of stop-words
- We used IDF to weigh dimensions but other collocation statistics can also be used
- Normalisation in cosine ignores the frequency effect, it normalised you the vector length, hence only the angle is important
- The corpus is crucial for the quality of relations
- Does cosine/semantic similarity always identify synonyms?
- Intrinsic and extrinsic evaluation
- Attributions and relational similarity: semantic relatedness vs similarity
- Word sense disambiguation is challenging
- Compositional vector-space models
- But what is a composition of sentence in terms of word contexts?
- Mitchell and Lapata compare various compositional operators, point wise multiplication works best
- Tensor product comes from Smolensky, it creates larger vector space
- Combinatory categorical grammar / CCG / Lambek pre-group grammar
- Grammar is defined as operations over semantic types: what vector representations correspond to the semantic types of the grammar?
- Tensor product is a multiplication of every dimension of the first space with every dimension of the second space, therefore the result is much larger, effectively comparing every dimension with every other dimension
- Verb is relational in nature, it relates a particular subject the particular

objects, hence the tensor product expresses these interactions
- Running example with plausibility vectors
- Baroni and Zamparelli, adjectives are arrived and nouns are vectors
- As composition is applied the semantic types are becoming smaller, i.e. chasing cat events
- How do we evaluate compositions? Inference

"Even taking account of laudable attempts such as Bos & Markert (2006), building on earlier work such as Alshawi (1992), it is probably fair to say that the classical logic-based enterprise in natural language processing has failed."

Interesting references

- Bruni, E., G. Boleda, M. Baroni, & N. Tran. 2012 (2012), Distributional semantics in technicolor, in Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, Jeju Island, Korea, (136–145).
- Pulman, Stephen (2012), Distributional semantic models, in Sadrzadeh Heunen & Grefenstette (eds.), Compositional Methods in Physics and Linguistics, Oxford University Press.

TFIDF, PMI, SVD

Chapter 6 Vector semantics and embeddings. D. Jurafsky and J. H. Martin. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. Third edition draft, Stanford University and University of Colorado at Boulder, December 30 2020.

M. Baroni. Composition in distributional semantics. Language and Linguistics Compass, 7(10):511–522, 2013.

From the class, VT23

- (Blue) Relations that cannot be captured:
  - metaphors, mapping between semantic domains, target and source
  - humour
  - irony
  - cultural knowledge
  - metaphors and literal words; X is an angel; love is a journey, X is going to explode; polysemy;
- Combining frameworks: distributional semantics, multi-modality, FrameNet
  - vector databases, knowledge graphs, edges
  - combining modalities; aligning (Purple)
  - mapping between formal and distirbutional representations; how the mapping is done? (Purple)
- (Green) Corpus dependence and types of representations
  - ethical issues, e.g. hate speech
  - checking data imposssible/hard

  - contexts in which words are used is reflected in documents, assuming a uniform language use
  - words not in a corpus are not represented
  - also an issue with language models (Orange), relation between vectors and ML models
- (Red) Large English data
  - communities where words are used differently
- (Purple), see above
- (Orange), see above
- Tensor product
  - [b, a] *[c, d]* –> [b*c, b*d, a*c, a*d]
  - [2, 3] [3,1] –> [2*3, 2*1, 3*3, 3*1] = [6, 2, 9, 3]
  - 2^n dimensions
  - see the linguistic example in the paper
- See on the differences between dot product, tensor product below
- (Grey) reference from the research paper (Logical negation is a challenge for distributional semantics, because predicates and their negations tend to occur in very similar contexts, and consequently their distributional vectors are very similar. Indeed, it is not even clear what properties a "negated" distributional vector should possess

From the class, VT22

- what kind of meaning do we represent
  - sense and reference?
  - lacks embodied and situated meaning, therefore does not "understand"
  - do models really need to "understand" embodied meaning
  - semantic relations - which ones and how to extend the model to represent different ones (e.g. adding polarity, morphology); hypernyms and meronyms, synonyms (e.g. run, rage and anger, annoyed we feel that they are synonyms but they are used in different contexts); words have different senses and these senses do not always overlap; we only have **one** semantic relation which is geometrically defined; how do we learn antonyms; there are no synonyms since each word would be used in relevant contexts
  - do words have "meaning" at all at any point: language is always in flux; "wrong" usages can become standard
- corpus matters
  - spoken, literature, synchronic and diachronic corpora, bias
  - issues with languages other than English
- computational challenges
  - the role of ML
  - computational benefits of modelling meaning this way
  - language generation from vectors (rather than understanding)
- applications for distributional representations
  - relations between sentences, not very well

- formal and distributional
  - complementary distribution: connecting with logic
  - distributional and logic representations and context: arguments and word contexts?
  - other formal grammars: CCG

From the class, VT21

- Groups 3, 4, 5, Max, 2, 1
- formal-distributional
- sense and reference: knowing a reference you cannot really know much about dogs and cars
- distance, vectors, words as atomic entities (subwords?)
- TF-IDF, IDF why inverse? why emphasise words in a few documents? dot product:
- TF-IDF: https://en.wikipedia.org/wiki/Tf%E2%80%93idf
- Dot product: https://en.wikipedia.org/wiki/Dot_product
- is there an alternative to TF-IDF: ML and PMI
- differences in meaning; semantic relations;
- are contexts meaningful: extra-linguistic features; are distributions really senses (Frege: sense allows you to identify reference); how much do we know about the individual from distributions
- does it make sense to compose vectors: interpretability of vectors themselves and of composition
- metaphors, performance
- similarity with human meaning comparison; how humans reason with meaning when they see a new word: world-knowledge associations
- lexical relations: how are different semantic relations reflected
- evaluation: intrinsic evaluation, precision at rank; second order context vector
- data?
- pattern-based approaches; word-net
- self-supervised models should be supervised by humans; interpretability
- challenges
- a lot of data, powerful machines; availability
- sparsity; open question
- type of datasets will give different vectors and representations (bias!)
- benefits
- easy to compare, can be extracted from corpus (no need for linguists!)
- applications
- search engines, thesauri
- compositionality
- phrases and sentences; what are features;
- extracting constituents from combined vectors
- the technicalities of composition?
- what representations can be used?

- does make sense to use the same compositionality for every word/constituent?
- similarity of sentences? or difference?
- concepts: CCG, formal grammars and pre-group grammars
- cancellation rule?
- tensor product and point-wise matrix multiplication
- point-wise matrix multiplication:
- interpretation in geometric space
- can relational types using ML;
- a . b = sum(a*i,b*i); sum the products of the position i in both vectors; gives us a single number/scalar; we reduce dimensions
- tensor product (symbol = circle with an x): this comes from a paper by Smolensky (1990), it's a matrix whose components are all the possible products u*iv*j of the components of vectors u and v; e.g. u1v1 u2v1 u3v1 for the first row and then for every u; Here we expand dimensions since we multiples two vectors and we've got a matrix
- point-wise matrix multiplication is multiplying each individual position of two matrices to create another matrix of the same dimension as the source matrices where the positions are products. (This is different than matrix multiplication).
- Hence, we have 3 different ways to combine information; we reduce dimensions, increase dimensions or keep the same dimensions. This will be relevant for modelling language with neural networks.

J. Mitchell and M. Lapata. Composition in distributional models of semantics. Cognitive Science, 34(8):1388–1429, 2010.