Vector semantics and embeddings unit

- Lexical semantics, words as vectors (**Today**)
- The Skipgram algorithm of Word2Vec (Wednesday)
- Begin transformers and machine translation (Friday)

Today:

- What vector semantics is and why we care
- The intuition of word-relatedness
- The idea of words as vectors

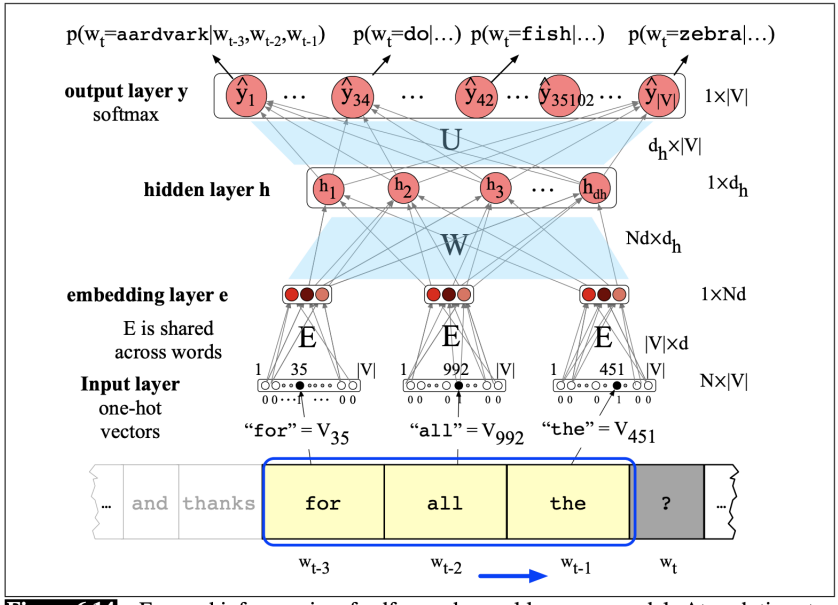

Figure 6.14 Forward inference in a feedforward neural language model. At each timestep

Jurafsky and Martin, Sec 6.3, pg 17

The position of a category boundary has psycholinguistic consequences. In an experimental task in which subjects have to say as quickly as possible whether or not an item belongs to a given category, it is found that the speed of response is slowest for items close to the category boundary. Hence, in (3), the ***bold*** items will be responded to slowest:

(3)

FRUIT	*apple*.....	***tomato***..... *potato*
VEHICLE	*car*.....	***bicycle***..... *chair*
WORD	*hand*.....	***malk***..... *pkhq*

The location of the boundary of a category is not determined by the location of its prototype core. Two categories may have more or less identical prototypes but their boundaries may be different; equally, two categories may have the same boundaries but differ in their prototypes. The first case can be illustrated by the French word *corde* and its English equivalent *rope*. Native informants suggest the same sort of examples as the best in the respective categories, from which we infer that the respective prototypes are at least very similar. However, there is evidence that the boundaries of CORDE and ROPE are not the same. The French dictionary *Le Petit Larousse* defines *ficelle* ("string") as "une corde mince" (= "a thin rope"); it would never occur to an English lexicographer to define *string* as "thin rope". It seems, therefore, that although FICELLE falls within the boundary of the category CORDE, STRING falls outside the boundary of the category ROPE, in other words, CORDE is the broader category. The converse case, of two categories with the same boundary but differing in the location of their cores, is harder to find, but perhaps COURAGE and BRAVERY are examples. There is something contradictory about *?The men showed great bravery, but no courage*. But there is evidence that their core regions are not the same. Student informants in an informal test rated (4) as a better example of a brave act, and (5) as a better example of a courageous act:

(4) A person jumps into a fast-flowing river in an attempt to save someone who has fallen in.

(5) A person risks his/her career and livelihood by exposing malpractice and injustice at the heart of government.

Alan Cruse, *Meaning in Language*, Oxford, 2011, pg 66

Vector semantics *is the standard way to represent word meaning in NLP . . . The roots of the model lie in the 1950s when two big ideas converged: [using] a point in three-dimensional space to represent the connotation of a word, and the proposal . . . to define the meaning of a word by its distribution in language use, meaning its neighboring words or grammatical environments.*

The idea of vector semantics is to represent a word as a point in a multi-dimensional semantic space that is derived from the distributions of word neighbors. Vectors for representing words are called **embeddings**.

Jurafsky and Martin, Chapter 6, pg 5 & 6

- **Valence.** The pleasantness of the stimulus
- **Arousal.** The intensity of emotion provoked by the stimulus
- **Dominance.** The degree of control exerted by the stimulus

	Valence	Arousal	Dominance
courageous	8.05	5.5	7.38
music	7.67	5.57	6.5
heartbreak	2.45	5.65	3.58
cub	6.71	3.95	4.24

Jurafsky and Martin, 6.1, pg 4 & 5

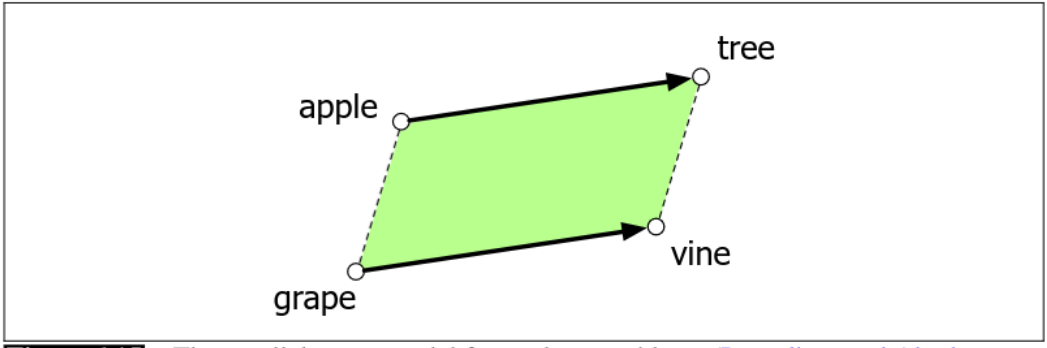

Figure 6.15 The parallelogram model for analogy problems (Rumelhart and Abrahamson, 1973): the location of $\overrightarrow{\text{vine}}$ can be found by subtracting $\overrightarrow{\text{apple}}$ from $\overrightarrow{\text{tree}}$ and adding $\overrightarrow{\text{grape}}$.

Coming up:

- Read from J&M Chapter 5 (Mon, Nov 17, **already done**)
- Read other part of J&M Chapter 5 (Wednesday, Nov 19)
- Work on stylo project (Fri, Dec 12)