Baroni & Zamparelli (2010): Nouns are vectors, adjectives are matrices

Representing adjective-noun constructions in semantic space

Min Fang

November 18, 2013

Compositionality (FS)

Different levels of "intersectivity"

- red car
- big mouse
- alleged thief
- fake Picasso

Insight of Formal Semantics:

- Attributive adjectives as *functions* from the meaning of the noun onto the meaning of a modified noun

Adjectives as linear maps I

- Mitchell&Lapata's additive model
  - $\mathbf{p} = \mathbf{A}\mathbf{u} + \mathbf{B}\mathbf{v}$

Baroni&Zamparelli's adaptation:

$$\mathbf{p} = \mathbf{B}\mathbf{v}$$

$\mathbf{p}_n$ . . . observed adjective-noun (AN) vector

$\mathbf{B}_{n \times n}$ . . . weight matrix representing a specific adjective

$\mathbf{v}_n$ . . . noun vector

Adjectives are linear functions from n-dimensional (noun) vectors onto n-dimensional vectors (expressed as matrix multiplication).

Adjectives as linear maps II

$$\mathbf{p} = \mathbf{B}\mathbf{v}$$

$$\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_n \end{pmatrix} = \begin{pmatrix} b_{1,1} & b_{1,2} & \dots & b_{1,n} \\ b_{2,1} & b_{2,2} & \dots & b_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n,1} & b_{n,2} & \dots & b_{n,n} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}$$

$$= \begin{pmatrix} b_{1,1}v_1 + b_{1,2}v_2 + \dots + b_{1,n}v_n \\ b_{2,1}v_1 + b_{2,2}v_2 + \dots + b_{2,n}v_n \\ \vdots \\ b_{n,1}v_1 + b_{n,2}v_2 + \dots + b_{n,n}v_n \end{pmatrix}$$

Adjectives as linear maps III

- estimate the values of **B** by partial least squares regression
  - independent variables: dimensions of the corpus-based vectors of the component nouns
  - dependent variables: dimensions of the corpus-observed AN vectors
- estimate a separate model for *each* adjective
  - vs. Guevara (2010): generic "AN-slot" function

Experimental Setup I

- Corpus
  - ukWaC corpus, mid-2009 dump of the English Wikipedia and the British National Corpus concatenated
- Vocabulary
  - AN test set: 26,440 ANs
    - contains 36 distinct adjectives chosen from various classes (734 ANs per adjective on average)
    - contains 1,420 distinct nouns
  - core vocabulary: 12K lemmas
    - top 8K noun lemmas
    - top 4K adjective lemmas
  - extended vocabulary: ~41K items
    - core vocabulary
    - test set ANs
    - 2.5K additional ANs

Experimental Setup II

- Semantic space
  - co-occurrence matrix: 41K×10K
    - columns: 10K lemmas that co-occur with the largest number of items in the core vocabulary
    - rows: all items in the extended vocabulary
  - dimensionality reduction
    - using Singular Value Decomposition (SVD)
    - 41K×10K $\Rightarrow$ 41K×300

1st Study I

Do the observed ANs correspond to our semantic intuition?

- Compute the centroids of observed AN vectors that share the same adjective
- Select the nearest neighbours (according to cosine similarity) in the extended vocabulary
- Position of the centroids
  - near the adjective itself
  - near the corresponding noun
  - near prototypical ANs for that adjective
  - near elements related to the definition of the adjective

black N	*historical N*	*easy N*	*necessary N*
black face	historical	easy start	necessary
black hand	hist. event	quick	necessary degree
black (noun)	hist. content	little cost	sufficient

1st Study II

- Select the nearest neighbours of specific ANs
- The neighbours can pick up the composite meanings

important route	*red cover*	*little war*	*historical map*
important transport	black cover	great war	topographical
important road	hardback	major war	atlas
major road	red label	small war	hist. material

Composition Methods I

- B&Z's adjective-specific linear map (*alm*)
  - $p_i = b_{i,1}\tilde{v}_1 + \cdots + b_{i,300}\tilde{v}_{300} + \epsilon_i$
  - solving 300 regression problems with 300 independent variables for each adjective
  - training data for each adjective ranges from 200 to >1K
  - PLSR, 50 latent variables
  - leave-one-out training regime

- Guevera's single linear map (*slm*)
  - concatenation of observed adjectives and noun vectors (independent variables)
  - coupled with corresponding AN vectors (dependent variables)
  - training data: 2K randomly sampled adjective-noun-AN tuples
  - estimation of a single coefficient matrix

Composition Methods II

- additive (*add*)
  - summing the corresponding adjective and noun vectors after normalisation
- multiplicative (*mult*)
  - component-wise multiplication of the adjective and noun vectors
- baselines
  - adjective (*adj*)
  - noun (*noun*)

2nd Study I

How well do the model-generated AN vectors approximate the unseen corpus-observed AN vectors?

- Compute for each generated AN vector the cosine similarity to the items in the extended vocabulary and rank the items accordingly
- Find the corpus-observed AN vector in the ranking

method	*25%*	*median*	*75%*
alm	17	170	$\geq$1K
add	27	257	$\geq$1K
noun	72	448	$\geq$1K
mult	279	$\geq$1K	$\geq$1K
slm	629	$\geq$1K	$\geq$1K
adj	$\geq$1K	$\geq$1K	$\geq$1K

2nd Study II

Problematic cases: anomaly of observed ANs?

SIMILAR		
adj N	*obs. neighbour*	*pred. neighbour*
common understanding	common approach	common vision
new actor	new cast	new cast
small droplet	droplet	drop
general question	general issue	general issue

DISSIMILAR		
adj N	*obs. neighbour*	*pred. neighbour*
American affair	Am. development	Am. policy
historical thing	different today	hist. reality
young photo	important song	young image
current dimension	left (noun)	current element

2nd Study III

- High inverse correlation between median rank and adjective frequency
  - *new* - 34, *great* - 79, *American* - 82, *large* - 82, *different* - 97
  - Model works best with frequent, highly polysemous adjectives
- High inverse correlation between the frequency of occurrences of an AN and the rank of the observed AN w.r.t. the predicted AN
  - Model is worse at approximating the observed vectors of rare forms

3rd Study I

Is it possible to compare adjectives under the assumption that they are functions?

- possible representations
  - *centroid*: centroid vector of the ANs that share an adjectiv
  - *matrix*: the estimated weight matrix of dimension 300×300 (unfolded into a 90K-dimensional vector)
  - *vector*: corpus co-occurrence profile vector of the adjective (projected onto the SVD-reduced space)
- evaluation: clustering task
  - adjective classes
    - colours: *white*, *black*, *red*, *green*
    - positive evaluation: *nice*, *excellent*, *important*, *major*, *appropriate*
    - time: *recent*, *new*, *current*, *old*, *young*
    - size: *big*, *huge*, *little*, *small*, *large*

3rd Study II

- clustering quality indicated by percentage purity
  - empirical 95% confidence intervals
  - baseline *random*: random assignments of adjectives to the clusters

input	*purity*
matrix	73.7 (68.4-94.7)
centroid	73.7 (63.2-94.7)
vector	68.4 (63.2-89.5)
random	45.9 (36.8-57.9)

Summary

- Adjective interpreted as a linear mapping of the noun vector
  - $\mathbf{p} = \mathbf{B}\mathbf{v}$
  - model-generated ANs should approximate corpus-observed ANs
  - estimate the values of the weight matrix by solving linear regression problems
- Experiments show that . . .
  - ANs in the corpus generally conform with our semantic intuitions and can be used as a goal of approximation (1st study)
  - B&Z's *alm* method provides the best approximation (2nd study), followed by the additive model
  - under the functional view, adjectives can still be meaningfully represented and compared even though the adjectives do not have a independently collected vector