Nouns are vectors, adjectives are matrices

статья by Marco Baroni & Roberto Zamparelli

презентед ту ю бай Гоша Лоншаков & Аня Полянская

Оглавление

прилагательные как функции

AN = A∩N? нет, см. *fake, different, good*

идея из формальной семантики

представление атрибутивных прилагательных как функция от значения существительного на значение модифицированного прилагательным существительного

суть

вклад исследования – новый метод дистрибутивной репрезентации AN, где прилагательное это линейная функция от вектора (noun representation) на другой вектор (AN representation).

модель - линейная регрессия

Adjectives as linear maps

p = Bv

p – observed AN vector
B – weight matrix representing the adjective at hand
v – noun vector

method: **partial least squares regression**

Experimental setup (corpus)

Concatenated corpus – сборник из ukWaC corpus, mid-2009 выгрузка English Wikipedia и the British National Corpus.

Tokenized, POS-tagged and lemmatized with the TreeTagger

2.83 billion tokens (excluding punctuation, digits, etc.)

Sample corpus – выборка из ukWaC corpus в 100M токенов

Experimental setup (vocabulary)

AN test set – 26,440 ANs (36 adjs x 1,420 nouns)

core vocabulary – top 8k nouns and 4k adjs most frequent lemmas from concatenated corpus (excluding top 50 most frequent nouns and adjs) 12k in total

extended vocabulary – core vocab. + AN test set + 16 adjs and 43 nouns + 2500 more ANs randomly sampled from sample corpus

In total, the **extended vocabulary** contains 40,999 entries: 8,043 nouns, 4,016 adjectives and 28,940 ANs

36 прилагательных

- size (*big, great, huge, large, major, small, little*)
- denominal (*American, European, national, mental, historical, electronic*),
- colors (*white, black, red, green*)
- positive evaluation (*nice, excellent, important, appropriate*)
- temporal (*old, recent, new, young, current*)
- modal (*necessary, possible*)
- some abstract antonymous pairs (*difficult, easy, good, bad, special, general, different, common*)

Experimental setup (semantic space construction)

Full co-occurrence matrix – top 10k lemmas (nouns, adjs and verbs) that co-occur with the largest number of items in the core set **(12k X 10k)**

Dimensionality reduction – using SVD (singular value decomposition)

12k X 10k ——–> 12k X 300

Experimental setup (composition methods)

adjective-specific linear map **(alm)**
n * **[A]** = AN

single linear map **(slm)** from previous work (Guevara 2010)
n * [] (+ a) = AN

additive **(add)** – adj vector + noun vector
norm(n) + norm(a) = AN

multiplicative **(mult)** – adj vector x noun vector
n * a = AN

Study 1: ANs in semantic space

Вычисление центроидов from normalized SVD space vectors всех AN, у которых одно и то же прилагательное (e.g. *American adult*, *American menu*, etc., summed to construct the *American N* centroid)

Центроид – среднее арифметическое векторов

Study 1: ANs in semantic space

American N	*black N*	*easy N*
Am. representative	black face	easy start
Am. territory	black hand	quick
Am. source	black (n)	little cost
green N	*historical N*	*mental N*
green (n)	historical	mental activity
red road	hist. event	mental experience
green colour	hist. content	mental energy
necessary N	*nice N*	*young N*
necessary	nice	youthful
necessary degree	good bit	young doctor
sufficient	nice break	young staff

Table 1: Nearest 3 neighbors of centroids of ANs that share the same adjective.

bad luck	*electronic communication*	*historical map*
bad	elec. storage	topographical
bad weekend	elec. transmission	atlas
good spirit	purpose	hist. material
important route	*nice girl*	*little war*
important transport	good girl	great war
important road	big girl	major war
major road	guy	small war
red cover	*special collection*	*young husband*
black cover	general collection	small son
hardback	small collection	small daughter
red label	archives	mistress

Table 2: Nearest 3 neighbors of specific ANs.

Study 2: Predicting AN vectors

applying methods

rank of observed vector in a list of cosine-ranked neighbors of predicted AN vector

and also compare to baseline vectors:

adj – adj vector

noun – noun vector

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

Table 3: Quartile ranks of observed ANs in cosine-ranked lists of predicted AN neighbors.

Study 2: Predicting AN vectors

SIMILAR			DISSIMILAR		
adj N	*obs. neighbor*	*pred. neighbor*	*adj N*	*obs. neighbor*	*pred. neighbor*
common understanding	common approach	common vision	American affair	Am. development	Am. policy
different authority	diff. objective	diff. description	current dimension	left (a)	current element
different partner	diff. organisation	diff. department	good complaint	current complaint	good beginning
general question	general issue	*same*	great field	excellent field	gr. distribution
historical introduction	hist. background	*same*	historical thing	different today	hist. reality
necessary qualification	nec. experience	*same*	important summer	summer	big holiday
new actor	new cast	*same*	large pass	historical region	large dimension
recent request	recent enquiry	*same*	special something	little animal	special thing
small drop	droplet	drop	white profile	chrome (n)	white show
young engineer	young designer	y. engineering	young photo	important song	young image

Table 4: Left: nearest neighbors of observed and *alm*-predicted ANs (excluding each other) for a random set of ANs where rank of observed w.r.t. predicted is 1. Right: nearest neighbors of predicted and observed ANs for random set where rank of observed w.r.t. predicted is $\geq 1K$.

Study 3: Comparing adjectives (how?)

300x300 матрицы развернуть в 90К вектора

1. color (white, black, red, green)
2. positive evaluation (nice, excellent, important, major, appropriate)
3. time (recent, new, current, old, young)
4. size (big, huge, little, small, large)

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)

Table 5: Percentage purity in adjective clustering with bootstrapped 95% confidence intervals.

Вопросы на подумать (лингвистического плана)

- критика выборки прилагательных
- почему *alm* works best with extremely frequent, highly polysemous adjectives like *new, large*, *different*?
- фразеологизмы? абсолютно некомпозициональне устойчивые синонимы-парафразы (aka *годовасик-тугоседя* as for *child* и другие продукты урбан-дикшонари)
- другие способы оценки?
- критика “классов прилагательных”?
- задел на сегментную морфологию (ака значения продуктивных аффиксов, в частности приставок)
- scale up? full sentences?

Вопросы на подумать (математического плана)

- размерности (as always)
- SVD
- Mutual Information