Applicative structure in vector space models

Márton Makrai, Dávid Nemeskey, András Kornai
Institute for Computer Science and Control, Hungarian Academy of Sciences

The problem: representing multiword expressions

Commutativity of vector addition is a fundamental feature of vector space models, e.g. Mikolov [3] suggests

$$\mathbf{v}\,(\text{king}) - \mathbf{v}\,(\text{queen}) = \mathbf{v}\,(\text{male}) - \mathbf{v}\,(\text{female})$$

By commutativity:

$\mathbf{v}\,(\text{king}) - \mathbf{v}\,(\text{male}) = \mathbf{v}\,(\text{queen}) - \mathbf{v}\,(\text{female})$ = 'ruler, gender unspecified'

If application of semantic functions is simply a vector to be added to the representation, the same logic would yield

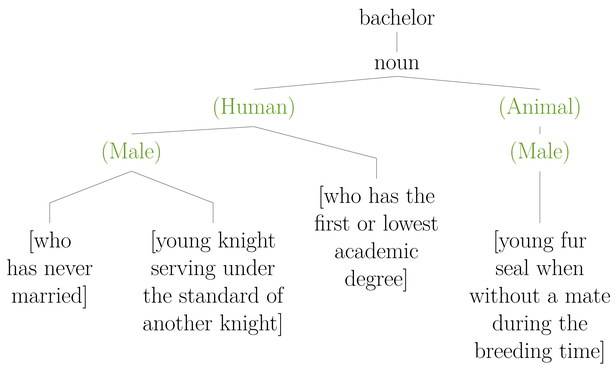
$$\mathbf{v}\,(\text{king of country})) = \mathbf{v}\,(\text{king}) + \mathbf{v}\,(\text{of}) + \mathbf{v}\,(\text{country}) =$$
$$= \mathbf{v}\,(\text{country}) + \mathbf{v}\,(\text{of}) + \mathbf{v}\,(\text{king}) = \mathbf{v}\,(\text{country of king})$$

Overview

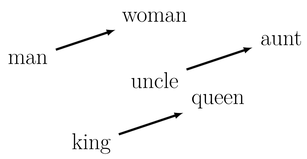
We introduce a new 100-dimensional embedding obtained by spectral clustering of a graph describing the conceptual structure of the lexicon. We use the embedding directly to investigate sets of antonymic pairs, and indirectly to solve the problem outlined above by treating ⊘ and ⊖ not as a vectors but as transformations.

Lexical decomposition

The standard model of lexical decomposition [2] divides lexical meaning in two parts, a systematic component, given by a tree of (generally binary) features, and an accidental component they call the [distinguisher].



Antonymic pair lists



For a set of male and female words, such as ⟨king, queen⟩, ⟨uncle,aunt⟩, ⟨actor, actress⟩, etc., the difference between words in each pair should represent the idea of gender. Similarly for pairs differing in some other feature. To test the hypothesis, we associated antonymic word pairs $\langle x_i, y_i\rangle$ from WordNet [4] to 26 classes e.g. END/BEGINNING, GOOD/BAD...

GOOD		VERTICAL		SIZE	
safe	out	raise	level	large	small
peace	war	tall	short	tall	short
pleasure	pain	rise	fall	expand	contract
ripe	green	north	south	loose	compact
defend	attack	shallow	deep	planar	linear
conserve	waste	ascending	descending	thicken	thin
affirmative	negative	superficial	profound	widen	narrow
⋮	⋮	⋮	⋮	⋮	⋮

Table 1: Word pairs associated to features

Test

- for k pairs $\mathbf{x}_i, \mathbf{y}_i$ we are looking for a common vector $\mathbf{a}$ such that
$$\mathbf{x_i} - \mathbf{y}_i \approx \mathbf{a}$$
- find $\text{argmin}_{\mathbf{a}}\, Err$
$$Err = \sum_i ||\mathbf{x_i} - \mathbf{y}_i - \mathbf{a}||^2$$
- $\text{argmin}_{\mathbf{a}}\, Err$ is actually the arithmetic mean of the vectors $\mathbf{x_i} - \mathbf{y}_i$
- is the minimal Err any better than what we could expect from a bunch of random $\mathbf{x_i}$ and $\mathbf{y}_i$?
- 100 random pairings of the words to estimate the error distribution, computing the minima of
$$Err_{rand} = \sum_i ||\mathbf{x_i}' - \mathbf{y}_i' - \mathbf{a}||^2$$
- is the error of the correct pairing, Err at least 2 or 3 standard deviations (σ) away from the mean of Err_{rand}?

Results with embeddings

# pairs	feature name	HLBL[5] original Err	m	σ	r	HLBL scaled Err	m	σ	r	SENNA[1] Err	m	σ	r	4lang Err	m	σ	r
42	vertical	1.77	2.62	0.0617	13.8	3.82	5.63	0.168	10.8	37.3	81.2	2.78	15.8	0.0672	0.1350	0.01360	4.98
156	good	1.92	2.29	0.032	11.6	4.15	4.94	0.0635	12.5	50.2	81.1	1.35	22.9	0.0589	0.0730	0.00218	6.45
32	many	1.56	2.46	0.0809	11.2	3.36	5.3	0.176	11	43.8	76.9	3.01	11	0.0516	0.0807	0.00681	4.26
49	in	1.94	2.62	0.0805	8.56	4.17	5.64	0.191	7.68	40.6	82.9	2.46	17.2	0.0553	0.0957	0.00551	7.33
65	active	1.87	2.27	0.0613	6.55	4.02	4.9	0.125	6.99	50.2	84.4	2.43	14.1	0.0790	0.0993	0.00553	3.68
28	end	1.68	2.49	0.124	6.52	3.62	5.34	0.321	5.36	34.7	76.7	4.53	9.25	0.0975	0.2430	0.03410	4.27
48	same	2.23	2.62	0.0684	5.63	4.82	5.64	0.14	5.84	49.1	80.8	2.85	11.1	0.0768	0.0976	0.00682	3.05
36	time	1.97	2.41	0.0929	4.66	4.26	5.2	0.179	5.26	51.4	82.9	3.06	10.3	0.0842	0.1210	0.00992	3.74
20	progress	1.34	1.71	0.0852	4.28	2.9	3.72	0.152	5.39	47.1	78.4	4.67	6.7	0.0676	0.0977	0.00847	3.56
34	yes	2.3	2.7	0.0998	4.03	4.96	5.82	0.24	3.6	59.4	86.8	3.36	8.17	0.0344	0.0726	0.00786	4.86
32	sophis	2.34	2.76	0.105	4.01	5.05	5.93	0.187	4.72	43.4	78.3	2.9	12	0.0665	0.0879	0.00858	2.50
12	color	1.2	1.59	0.104	3.7	2.59	3.47	0.236	3.69	46.1	70	5.91	4.04	0.0564	0.0681	0.01940	0.600
18	mental	1.86	2.14	0.0783	3.54	4.02	4.6	0.155	3.76	51.9	73.9	3.52	6.26	0.0486	0.0601	0.00329	3.51
23	whole	1.96	2.19	0.0718	3.2	4.23	4.71	0.179	2.66	52.8	80.3	3.18	8.65	0.0996	0.2000	0.02120	4.74
14	gender	1.27	1.68	0.126	3.2	2.74	3.66	0.261	3.5	19.8	57.4	5.88	6.38	0.0820	0.2830	0.05330	3.76
17	strong	1.41	1.69	0.0948	2.92	3.05	3.63	0.235	2.48	49.5	74.9	3.34	7.59	0.0693	0.0686	0.01111	0.0625
16	know	1.79	2.07	0.0983	2.88	3.86	4.52	0.224	2.94	47.6	74.2	4.29	6.21	0.0598	0.0794	0.00706	2.77
12	front	1.48	1.95	0.17	2.74	3.19	4.21	0.401	2.54	37.1	63.7	5.09	5.23	0.0551	0.0756	0.01020	2.01
22	size	2.13	2.69	0.266	2.11	4.6	5.86	0.62	2.04	45.9	73.2	4.39	6.21	0.0299	0.0452	0.00514	2.98
10	distance	1.6	1.76	0.0748	2.06	3.45	3.77	0.172	1.85	47.2	73.3	4.67	5.58	0.0353	0.0351	0.00456	0.0438
10	real	1.45	1.61	0.092	1.78	3.11	3.51	0.182	2.19	44.2	64.2	5.52	3.63	0.0638	0.0920	0.01420	1.98
8	sound	1.65	1.8	0.109	1.36	3.57	3.88	0.228	1.37	46.2	62.7	6.17	2.67	0.0565	0.0656	0.01830	0.495
14	primary	2.22	2.43	0.154	1.36	4.78	5.26	0.357	1.35	59.4	80.9	4.3	5	0.0890	0.0895	0.00928	0.0529
8	single	1.57	1.82	0.19	1.32	3.38	3.83	0.32	1.4	40.3	70.7	6.48	4.69	0.0450	0.0833	0.01970	1.95
7	hard	1.46	1.58	0.129	0.931	3.15	3.41	0.306	0.861	42.5	60.4	8.21	2.18	0.0312	0.0521	0.01960	1.06
10	angular	2.34	2.45	0.203	0.501	5.05	5.22	0.395	0.432	46.3	60	6.18	2.2	0.0323	0.0363	0.00402	0.999

Table 2: Error of approximating real antonymic pairs (Err), mean and standard deviation (m, σ) of error with 100 random pairings, and the ratio $r = \frac{|Err - m|}{\sigma}$ for different features and embeddings

Results with embeddings (conted)

In Table 2,

- features above the first line → antonymic relations are well captured by the embeddings
- features below the second line → antonymic relations are not captured by the embeddings
- the more pairs, the better result

Embedding based on conceptual representation

- Input: a graph
  - nodes are concepts
  - $A \to B$ iff B is used in the definition of A
- base vectors are obtained by the spectral clustering method pioneered by [6]:
  - the incidence matrix of the conceptual network is replaced by an affinity matrix whose ij-th element is formed by computing the cosine distance of the ith and jth row of the original matrix, and
  - the first few (in our case, 100) eigenvectors are used as a basis.
- a word w_i in the basic vocabulary is included in the graph and corresponds to a base vector b_i
- for other words w in the dictionary, we take the definition of any word w in the Longman Dictionary of Contemporary English, we form $V(w)$ as the sum of the b_i for the w_is that appeared in the definition of w (with multiplicity)
- stopwords: the 19 most frequent words

HLBL and SENNA vs 4lang

Judgments under the three given embeddings and 4lang are highly correlated, see table 3. Unsurprisingly, the strongest correlation is between the original and the scaled HLBL results. Both the original and the scaled HLBL correlate notably better with 4lang than with SENNA, making the latter the odd one out.

	HLBL original	HLBL scaled	SENNA	4lang
HLBL original	1	0.925	0.422	0.856
HLBL scaled		1	0.390	0.772
SENNA			1	0.361
4lang				1

Table 3: Correlations between judgments ($r \overset{?}{\leqslant} 2\sigma$) based on different embeddings

Applicative structure

- the dictionary-based embedding enables us to investigate the function application issue
- asymmetric expressions: john HAS dog, dog HAS john
- 4lang: a semantic representation in which predicates have at most two arguments
- two transformations T_1 and T_2 corresponding to argument (indice)s regulate the linking of arguments:
  James kills — *James* is agent — $\mathbf{v}\,(\text{James}) + T_1\,(\mathbf{v}\,(\text{kill}))$
  kills James — *James* is patient — $T_2\,(\mathbf{v}\,(\text{kill})) + \mathbf{v}\,(\text{James})$
- distinguish also agent and patient *relatives* as in
  - *the man that killed James* versus
  - *the man that James killed.*

References

[1] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research (JMLR)*, 2011.

[2] J. Katz and Jerry A. Fodor. The structure of a semantic theory. *Language*, 39:170–210, 1963.

[3] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2013)*, pages 746–751, Atlanta, Georgia, 2013. Association for Computational Linguistics.

[4] George A. Miller. Wordnet: a lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.

[5] Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. *Advances in neural information processing systems*, 21:1081–1088, 2009.

[6] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856. MIT Press, 2001.

Acknowledgments

Work supported by OTKA grant #82333.

MTA Hungarian Academy of Sciences
SZTAKI Institute for Computer Science and Control