Applicative structure in vector space models

Márton Makrai, Dávid Nemeskey, András Kornai

HAS Computer and Automation Research Institute

The problem

Mikolov [3] suggests

$$\textbf{king} - \textbf{queen} = \textbf{male} - \textbf{female}$$

By commutativity:

$\textbf{king} - \textbf{male} = \textbf{queen} - \textbf{female} =$ ‘ruler, gender unspecified’

But with function application:

$$\textbf{Victoria} = \textbf{queen} \oslash \textbf{England} \quad \text{and}$$
$$\textbf{Victor} = \textbf{king} \oslash \textbf{Italy}$$

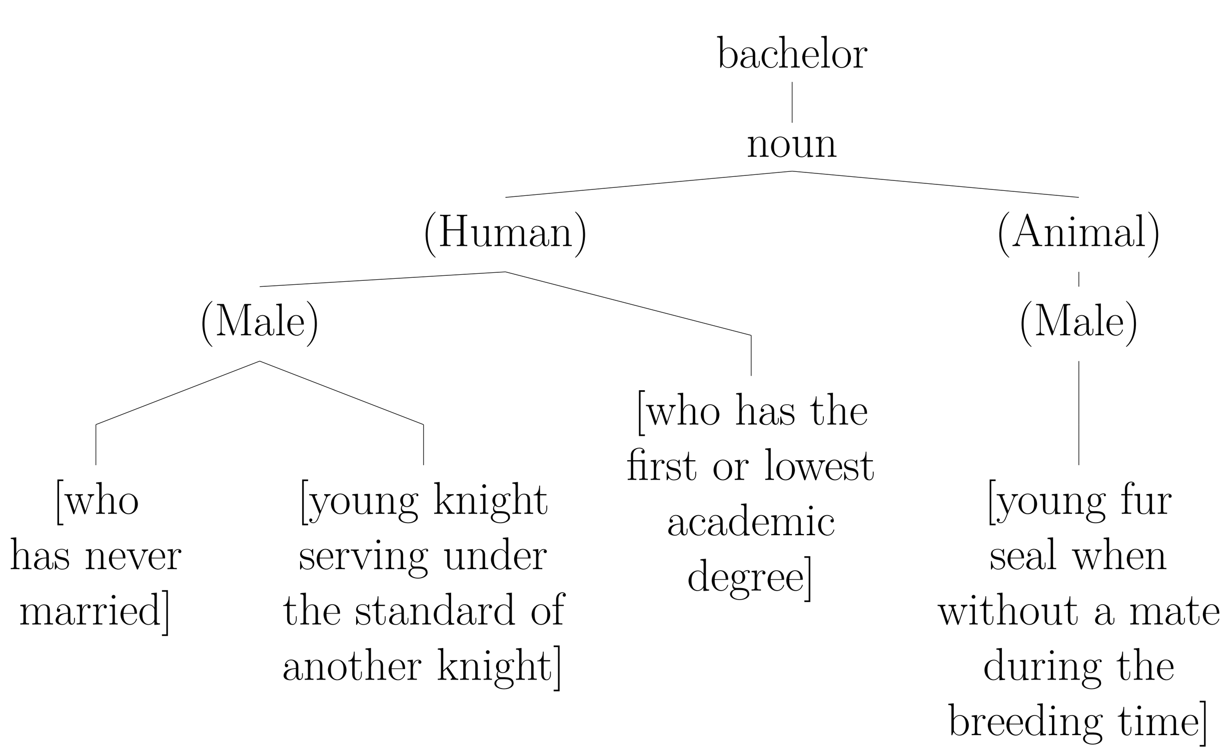
If the function application operator $\oslash$ is simply another vector to be added to the representation, the same logic would yield that Italy is the male counterpart of female England.

Overview

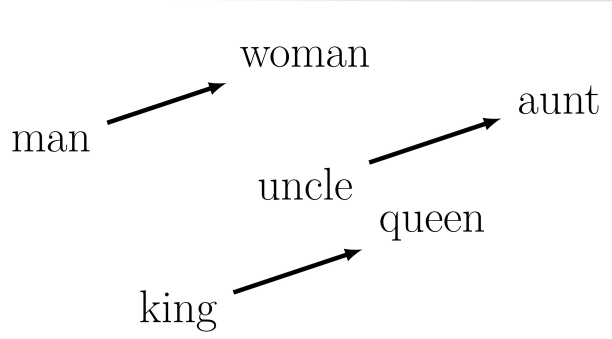
We introduce a new 100-dimensional embedding obtained by spectral clustering of a graph describing the conceptual structure of the lexicon. We use the embedding directly to investigate sets of antonymic pairs, and indirectly to solve the problem outlined above by treating $\oslash$ and $\otimes$ not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

Lexical decomposition

The standard model of lexical decomposition [2] divides lexical meaning in 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	
safe	out	raise	level
peace	war	tall	short
pleasure	pain	rise	fall
ripe	green	north	south
defend	attack	shallow	deep
conserve	waste	ascending	descending
affirmative	negative	superficial	profound
⋮	⋮	⋮	⋮

Table 1: Word pairs associated to features GOOD and VERTICAL

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}?

- 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
- caused by size?

Embedding based on conceptual representation

- Input: a graph
  - nodes are concepts
  - $A \rightarrow B$ 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

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
32	many	40.5	65.8	2.69	9.39	40.5	65.8	2.82	8.98	1.27e+03	2.28e+03	96.4	10.5	0.627	0.789	0.077	2.11
42	vertical	69.1	99.1	3.43	8.74	69.1	98.9	3.58	8.34	1.38e+03	2.94e+03	122	12.8	0.808	1.69	0.203	4.34
156	good	254	301	6.74	6.96	254	302	6.19	7.79	6.47e+03	1.05e+04	229	17.5	3.78	4.38	0.186	3.26
49	in	69.7	94.9	4.27	5.92	69.7	94.5	4.35	5.7	1.71e+03	3.55e+03	128	14.4	1.13	1.63	0.137	3.68
48	same	93.8	112	3.29	5.48	93.8	113	3.11	6.04	2.11e+03	3.53e+03	120	11.8	1.39	1.71	0.149	2.09
20	progress	21.5	28.5	1.56	4.45	21.5	28.7	1.44	5	801	1.37e+03	86.2	6.62	0.432	0.67	0.0679	3.5
28	end	35.3	51.8	3.75	4.41	35.3	52.8	3.67	4.79	798	1.78e+03	137	7.14	0.748	3.6	0.539	5.3
12	color	10.8	14.6	0.978	3.9	10.8	14.8	1.09	3.67	461	709	72.8	3.4	0.171	0.155	0.0493	0.319
18	mental	31.7	36.2	1.31	3.45	31.7	36.3	1.14	4.08	830	1.2e+03	57.4	6.4	0.605	0.694	0.0596	1.49
65	active	95.2	112	5.19	3.32	95.2	113	5.36	3.32	2.51e+03	4.07e+03	196	7.96	1.75	1.95	0.142	1.45
36	time	59.2	70.4	3.43	3.26	59.2	70	3.42	3.16	1.49e+03	2.36e+03	113	7.68	0.845	1.46	0.175	3.5
32	sophis	65.6	74.7	2.84	3.21	65.6	75.4	2.86	3.42	1.26e+03	2.25e+03	93.3	10.6	0.864	0.988	0.106	1.17
23	whole	39.3	45.1	1.87	3.14	39.3	45.4	1.91	3.21	1.06e+03	1.65e+03	84.1	7.07	0.706	1.4	0.216	3.19
34	yes	62.1	70.8	3.45	2.52	62.1	70.6	3.84	2.22	1.54e+03	2.29e+03	122	6.12	0.306	0.703	0.137	2.89
12	front	11.9	16.5	2.15	2.14	11.9	16.1	2.25	1.87	371	635	73.8	3.58	0.201	0.26	0.0539	1.1
8	single	7.85	10.4	1.31	1.94	7.85	10.4	1.54	1.64	282	529	56.1	4.41	0.107	0.166	0.0516	1.15
14	primary	24.4	28.1	2.15	1.74	24.4	28.4	1.99	2	713	1.01e+03	85.3	3.47	0.547	0.505	0.0583	0.718
14	gender	15.3	18.3	1.88	1.62	15.3	18.3	1.74	1.73	258	655	70.6	5.62	0.5	2.51	0.497	4.04
8	sound	11.6	12.7	0.744	1.52	11.6	12.7	0.833	1.32	324	444	44.8	2.68	0.138	0.142	0.0397	0.112
16	know	25.1	27.2	1.83	1.18	25.1	27.2	1.93	1.09	714	1.04e+03	65.3	5.02	0.435	0.611	0.0766	2.29
10	angular	18.8	16.3	2.19	1.14	18.8	16.3	2.03	1.22	371	457	49.9	1.73	0.158	0.16	0.0288	0.0757
10	real	13	13.9	1.09	0.808	13	14	1.13	0.844	442	612	54	3.15	0.223	0.286	0.0555	1.14
10	distance	16	16.7	1.05	0.676	16	16.7	1.15	0.577	472	706	66.1	3.53	0.109	0.0799	0.0172	1.69
17	strong	21.2	22.2	1.54	0.615	21.2	22.1	1.59	0.583	693	911	68.6	3.18	0.596	0.446	0.108	1.39
22	size	44.8	45.3	5.88	0.0856	44.8	45.9	5.45	0.211	1.01e+03	1.36e+03	127	2.74	0.27	0.314	0.0474	0.929

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

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	0.921	0.25	0.458
HLBL scaled		1	0.23	0.529
SENNA			1	0.196
4lang				1

Table 3: Correlations between judgments based on different embeddings

Application

- 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 to regulate the linking of arguments

James kills	*James* is agent	$V(\text{James}) + T_1 V(\text{kill})$
kills James	*James* is patient	$V(\text{James}) + T_2 V(\text{kill})$

- distinguish 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 Zweig Geoffrey. Linguistic regularities in continuous spaceword representations. In *Proceedings of NAACL-HLT-2013*, 2013.

[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 SZTAKI
COMPUTER AND AUTOMATION RESEARCH INSTITUTE
HUNGARIAN ACADEMY OF SCIENCES