Denoising composition in distributional semantics

Gábor Borbély
Department of Algebra
Budapest University of Technology
Egry József u. 1
1111 Budapest, Hungary
borbely@math.bme.hu

András Kornai
Institute for Computer Science
Hungarian Academy of Sciences
Kende u. 13-17
1111 Budapest, Hungary
andras@kornai.com

Marcus Kracht
Department of Linguistics
Bielefeld University
Universitätsstrasse 25
33615 Bielefeld, Germany
marcus.kracht@uni-bielefeld.de

Dávid Nemeskey
Institute for Computer Science
Hungarian Academy of Sciences
Kende u. 13-17
1111 Budapest, Hungary
nemeskeyd@sztaki.mta.hu

1 Introduction

Since in distributional semantics we derive the embedding from a corpus, and the corpus is just a sample from the entire distribution, it is inevitable that the vectors obtained in the process will be somewhat noisy. In Section 2 we analyze this and other sources of noise, and in Section 3 we turn to the question of how much the considerations of compositionality discussed in Kornai and Kracht (2015) are affected by noise.

We begin with some simple analytic considerations. In d dimensions, the unit ball has surface area $2\pi^{d/2}/\Gamma(d/2)$. If we equally divide this area among n cones, each peaking at the origin and having half angle θ, the surface area cut out by one cone is equal to $1/n$th of the total surface:

$$\frac{2\pi^{\frac{d-1}{2}}}{\Gamma(\frac{d-1}{2})}\int_0^\theta \sin^{d-2}\phi\,\mathrm{d}\phi = \frac{1}{n}\cdot\frac{\pi^{d/2}}{\Gamma(d/2)} \qquad (1)$$

The first point of interest is $n \sim 10^5, d = 300$, giving us a noise cone for GloVe (Pennington, Socher, and Manning, 2014) cosine similarity of about 0.25. It has been observed by (Mikolov, Yih, and Zweig, 2013) that in an analogy $a : b = c : d$ we can calculate v_d approximately as $v_a - v_b + v_c$ or, what is the same, we expect $v_a - v_b = v_c - v_d$. Since that time, methods other than straight addition were found to be slightly more efficient for solving analogy tasks (Levy and Goldberg, 2014), but here we retain the additive framework for ease of calculation. In practice, the winners in analogy task lookups often display cosine similarities in the 0.4-0.5 range, well above the noise level.

2 Noise effects

For empirical data, we took a standard English corpus, the UMBC Webbase (Han et al., 2013), and a new Hungarian corpus of comparable size. We cut UMBC in two roughly equal parts in two ways. In the even-odd cut (top panel in Table 1) we formed the two subcorpora by alternating between paragraphs (even-odd). In the begin-end cut (mid panel) we chose the first and the second half of UMBC. The bottom panel again shows the even-odd cut, but this time for GloVe vectors trained on a morphologically analyzed Hungarian corpus where the stem was treated as separate from the suffix (see Section 3). After running separately GloVe on the odd and the even parts, we compared the cosine similarities of the vectors in the two embeddings by five different methods, and we repeated the experiment comparing the beginning and end halves of UMBC, and the even-odd cut on the Hungarian corpus.

Since the first (direct) comparison method shows basically uncorrelated vectors (first line in all panels), we need to look at stability by bringing the vectors obtained from the two subcorpora into closer alignment. We do this by using the best orthonormal transformation (rot), or just by the best general linear transformation (gl), without normalization for vector length (nolen) and with normalization (len). Lines 2-5 of all panels show the results for the first 100 most frequent words (column @100); of 100 less frequent words, those ranked between 4,900 and 5,000 (column @5k); and the average similarity of the first 50k words (column 50k).

cut	cond	@100	@5k	50k
even-odd	direct	.010	.004	.003
even-odd	nolen-rot	.973	.946	.863
even-odd	len-rot	.973	.945	.862
even-odd	nolen-gl	.977	.955	.880
even-odd	len-gl	.976	.952	.879
beg-end	direct	.002	.004	.003
beg-end	nolen-rot	.966	.898	.764
beg-end	len-rot	.966	.897	.763
beg-end	nolen-gl	.965	.908	.789
beg-end	len-gl	.964	.903	.787
Hun e-o	direct	.357	.107	.072
Hun e-o	nolen-rot	.905	.884	.824
Hun e-o	len-rot	.903	.881	.823
Hun e-o	nolen-gl	.908	.899	.846
Hun e-o	len-gl	.903	.894	.844

Table 1: Paired cosine similarities

As expected, general linear transforms always perform better than rotations, and more frequent words show more stability than the average, since they are better trained. Length normalization has very little impact. In fact the top panel of similarities are good enough to conclude that swapping one set of vectors against the other will not affect the performance of the system on the Google analogy (GA) and similar tasks by much, and this is borne out by the facts: with the vectors trained on the even paragraphs we obtain 71.5%, and with those trained on the odd paragraphs 71.7%, much as expected.

The effects of the beginning-end split are 6 times bigger: 71.8% v. 70.7%, and we may even consider the first half of UMBC a separate corpus from the second (Curran and Osborne, 2002). To the extent GA probes the additive structure of the semantic space, it is fair to conclude that a minor shift in the training data, such as seen in interleaved halves of the same corpus, leaves the additive structure intact, but using more disjoint corpora, even if collected by the exact same methods, actually affects this structure.

In our next set of experiments (top panel of Table 2) we considered sparse overcomplete representations computed of the same GloVe vectors by the method of Faruqui et al. (2015)[1]. Since the direct comparison is useless, and length normalization makes very little difference, we only show the rotated and general linear similarities. It should be understood that the raw numbers are not directly comparable to those obtained for GloVe, since here $d = 3000$, for which Eq. (1) yields .08, about a third of the .25 we obtained in 300 dimensions. In this light, we actually see the sparse vectors as *more* stable than the raw GloVe vectors we obtained them from.

We also considered the nonnegative vectors in the same spirit as Faruqui et al. (2015) (middle panel of Table 2) and these are again quite stable. In fact, their stability is quite remarkable if we take into account the fact that everything now must be crammed in the first hyperquadrant. Yet this sparsification process loses a great deal of the linear analogical structure that was present in the GloVe vectors, and the resulting set of vectors now perform only at the 44.2% (even training) or 45.7% (odd training) level on GA.

vecs	dim	cond	@100	@5k	50k
Sparse	3k	nolen-rot	.627	.536	.458
Sparse	3k	nolen-gl	.754	.688	.600
Nonneg	3k	nolen-rot	.532	.477	.415
Nonneg	3k	nolen-gl	.621	.599	.553
k=5	2k	nolen-rot	.523	.466	.505
k=5	2k	nolen-gl	.583	.515	.561

Table 2: Paired cos similarities (sparse vectors)

We also investigated the considerably more sparse vectors suggested by Arora et al. (2016), reimplementing their method using the `pyksvd` library[2] (bottom panel of Table 2). These provide even better stability results. As the last block of Table 1 shows, the average cosine similarity over the most frequent 50k words is close to those obtained for the earlier sparse vector, even though these have at most 5 nonzero components out of 2,000 in contrast to the earlier ones which had 3-600 nonzero elements out of 3,000. Here the noise cone has cosine half-angle 0.095, about 20% larger than in the previous conditions. That the results are extremely stable is further driven home by the fact that when we compare (by linear transform) either set of our vectors to their original embedding (data kindly provided by Sanjeev Arora and Yinhyu Liang), the alignment is very similar, 0.583, even though they trained on a different corpus, the English Wikipedia. By this time, however, even more of the linear structure is lost, as these embeddings achieve only 23.1% (even training) and 23.5% (odd training) on GA.

It is not at all surprising that the most frequent 100 words (@100 column in both Tables 1 and

[1] https://github.com/mfaruqui/sparse-coding

[2] https://github.com/hoytak/pyksvd

2) show markedly better stability than those between frequency ranks 4,900 to 5,000 (@5k column), and the expectation would be that averaging the cosine distances for the first 50k words (last column) would. Yet the sparsest vectors (bottom panel of Table 2) do not show this effect, which indicates that k-SVD is somehow closer to the latent semantic structure or, at the very least, it is less sensitive to data frequency. On the Simlex-999 word similarity task (Hill, Reichart, and Korhonen, 2014), the 3k-dimensional nonnegative sparse vectors obtained by the method of Faruqui et al. (2015) slightly overperform the standard GloVe vectors, with a Spearman ρ of .395 (even) and .382 (odd) where the baseline has .371. The sparser vectors computed by the method of Arora et al. (2016) are less effective, with $\rho = .312$ (even) and .289 (odd). Overall, we see better stability at the expense of less preserved similarity and analogy structure, while in an ideal world the two should go hand in hand. The next Section formulates a preliminary hypothesis why we see this pattern, based on Hungarian data.

3 Compositionality

Here we consider the space V to be decomposed into two largely orthogonal subspaces G ("grammatical") and M ("meaningful"), roughly corresponding to the traditional distinction between function and content words. In Kornai and Kracht (2015) we used the compositional mechanism of Context Vector Grammars (Socher et al., 2013) to demonstrate that grammatical formatives such as the deadjectival adverb-forming suffix *-ly* or the comparative *-er* must contribute additively to the representations, so that e.g. $\vec{bigger} = \vec{big} + \vec{er}$. Obviously, if this holds every 'grammatical' GA task such as `gram1-adjective-to-adverb` or `gram5-present-participle` is solved automatically. Here we explore the overall hypothesis, derived from first principles, that paradigmatic contrasts between two forms of the same stem yield a vector that lies entirely in G, so that e.g. all plural forms must be obtained by addition of a single $\vec{pl} \in G$ to the vector of the singular (unmarked) stem in M.

To study stems and inflections separately, and investigate the effects of main category (noun, verb, adjective, adverb) on both G and M, in laboratory pure form, we took a large corpus of a highly agglutinative language, Hungarian, and by morphological analysis produced a de-glutinized version where the stem and the paradigmatic suffixes are separated by a whitespace the same way two words would. Perhaps inevitably, this became known as the 'gluten free' corpus of Hungarian, GFH. Simple search for neighboring words shows GFH to be very strongly clustered: for example the nearest neighbors of Hungarian male first names like *András* are *Imre, Lajos, József, István, Károly, Sándor, László, Péter, Géza*, while the 10 nearest neighbors of *Marcus* are *Kevin, Phil, Ian, David, Brian, Gary, Jeff, Jason, Matthew*. Remarkably, postpositions such as *alatt* cluster not just with other postpositions but also with case endings: the nearest neighbors are the terminative, inessive, superssive cases, the postposition *után* 'after', the adessive case, the postposition *között* 'between', followed by the illative, sublative, and instrumental cases.

Equation (1) of Arora et al. (2016) states that Pr(w occurs in discourse c) is proportional to $\exp(c \cdot v_w)$, but the reasoning behind this holds only for vectors from M. We expect Pr(w occurs in discourse c) to be constant for members of G (different constants for different function words, of course), owing to the fact that they are topic independent. In terms of grammatical formatives combining with stems, we expect no surprises in case all meanings share the same grammatical class: only adjectives can take comparative and superlative forms, only nouns will take case endings, only verbs will take tensed (finite) and infinitive forms, etc. To the extent different word senses co-occur with different syntactic environments, we can invoke more complex estimation methods that assign different vectors to different senses (Huang et al., 2012; Neelakantan et al., 2014; Bartunov et al., 2015), but for now we concentrate only on those stems that have only one part of speech, by filtering out the rest.

Clustering an English embedding, such as GloVe/UMBC used above, for part of speech, is not a trivial task. When we use only those words that have a unique Penn tag, and look at the average vector belonging to that tag, what we find that the variance of this vector is generally larger than the distance separating two centroids. This is depicted in Figure 1, which shows the centroids projected down to two dimensions together with their one sigma radius environment which is turned by the projection into a polygon. Different colors cor-

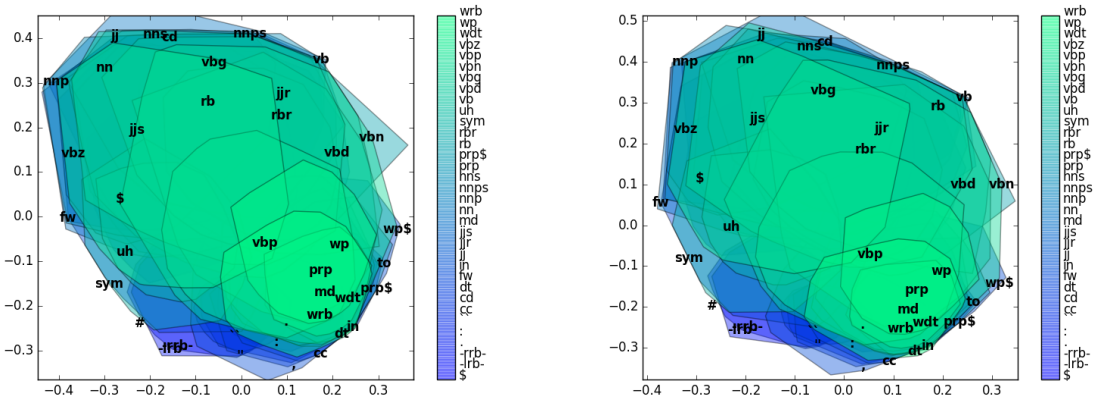

Figure 1: POS centroids based on UMBC odd and rotated even

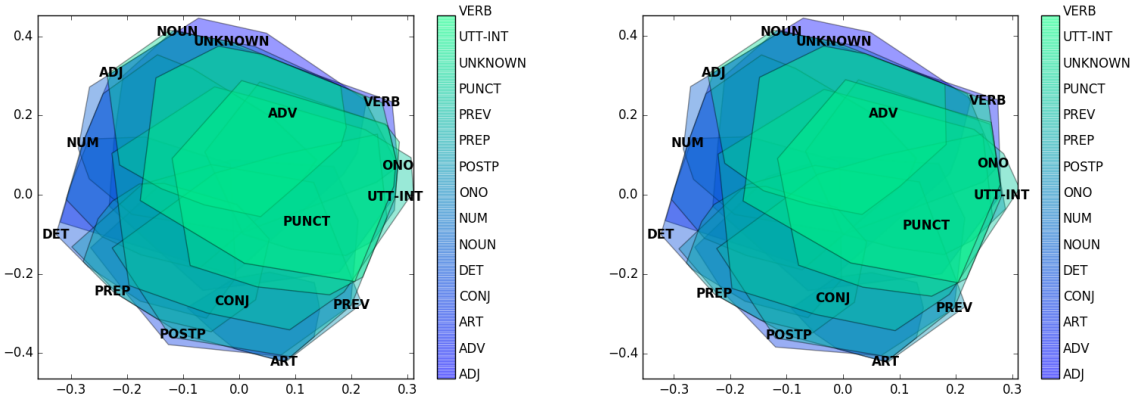

Figure 2: POS centroids based on GFH odd and rotated even

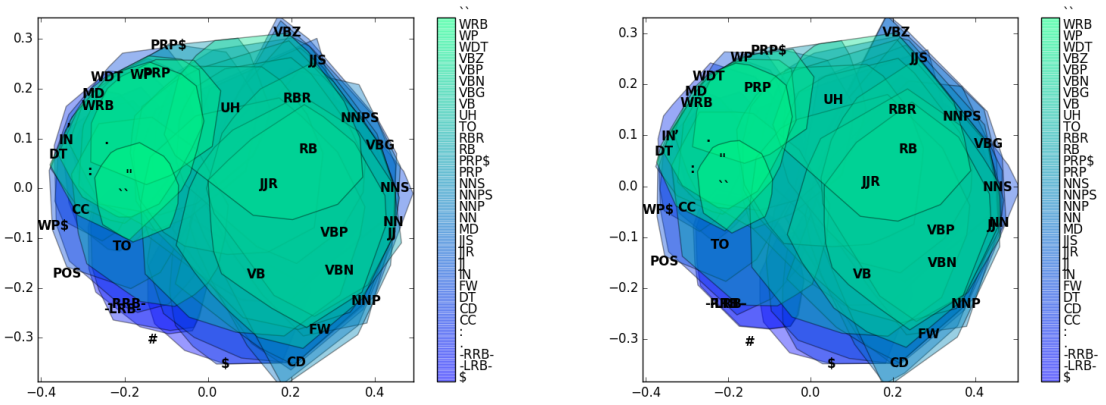

Figure 3: POS centroids based on GF-UMBC odd and rotated even

respond to the different parts of speech.

It is here that a careful study of noise begins to pay off: the left and right panels of Fig. 1 are quite different, yet they were obtained from what is, for all intents and purposes the same corpus, UMBC-odd and UMBC-even (rotated for best fit with odd). What we want instead is a stable picture, quite independent of which half of the corpus it is based on, and showing good separation between the categories. This is what we obtain in

GFH (see Fig 2). Note that the content categories are all separated from one another, and cluster near the top, while the function categories (except for UNK, unknown word, which is really a content category) cluster at the bottom.

As Fig. 3 shows, the "gluten free" version of UMBC, obtained by running the Stanford Parser's morphology analysis, is considerably better than the original, and by systematic analysis of the Penn tagset we should be able to obtain a geometrically just as clear picture as for Hungarian.

4 Summary and conclusions

We argued that the geometry of embeddings is much easier to understand than one could see based on raw data: major categories (N, V, Adv, A) fall in their own clusters, and so do function words (whose clusters are much smaller). Further, it is reasonable to conceive of the entire space as being the direct sum of the grammatical subspace G and the meaning subspace M. The two follow different probability laws: function word probabilities are independent of discourse-level context, content word probabilities follow Arora's Eq. 1. In future work, we hope to demonstrate the same conclusions for the sparse overcomplete and the atomic embeddings of Faruqui et al. (2015) and Arora et al. (2016) as well.

References

Arora, Sanjeev et al. (2016). "Linear Algebraic Structure of Word Senses, with Applications to Polysemy". In: *arXiv:1601.03764v1*.

Bartunov, Sergey et al. (2015). "Breaking Sticks and Ambiguities with Adaptive Skip-gram". In: *ArXiv preprint*.

Curran, James R. and Miles Osborne (2002). *A very very large corpus doesn't always yield reliable estimates*.

Faruqui, Manaal et al. (2015). "Retrofitting Word Vectors to Semantic Lexicons". In: *Proceedings of NAACL 2015*. Best Student Paper Award.

Han, Lushan et al. (2013). "UMBC_EBIQUITY-CORE: Semantic textual similarity systems". In: *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pp. 44–52.

Hill, Felix, Roi Reichart, and Anna Korhonen (2014). "Multi-modal models for concrete and abstract concept meaning". In: *Transactions of the Association for Computational Linguistics* 2.10, pp. 285–296.

Huang, Eric H. et al. (2012). "Improving Word Representations via Global Context and Multiple Word Prototypes". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*. ACL '12. Jeju Island, Korea: Association for Computational Linguistics, pp. 873–882. URL: http://dl.acm.org/citation.cfm?id=2390524.2390645.

Kornai, András and Marcus Kracht (2015). "Lexical Semantics and Model Theory: Together at Last?" In: *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 14)*. Chicago, IL: Association for Computational Linguistics, pp. 51–61.

Levy, Omer and Yoav Goldberg (2014). "Linguistic Regularities in Sparse and Explicit Word Representations". In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*. Ann Arbor, Michigan: Association for Computational Linguistics, pp. 171–180. URL: http://aclweb.org/anthology/W14-1618.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig (2013). "Linguistic Regularities in Continuous Space Word Representations". In: *Proceedings of NAACL-HLT 2013*, pp. 746–751.

Neelakantan, Arvind et al. (2014). "Efficient Non-parametric Estimation of Multiple Embeddings per Word in Vector Space". In: *EMNLP*.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global Vectors for Word Representation". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Socher, Richard et al. (2013). "Parsing with compositional vector grammars". In: *The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*.