A view of LSA/ESA in Computational Psychoanalysis

Giuseppe Iurato

A view of LSA/ESA in Computational Psychoanalysis

Giuseppe Iurato*

University of Palermo, Palermo, Italy.

Abstract

We wish simply to highlight a possible conceptual parallelism between some key points emerging from the pattern comparison of Latent Semantic Analysis (LSA) and Explicit Semantic Analysis (ESA), with the Freudian method of free associations.

Latent semantic analysis (in short, LSA) is an approach to automatic indexing and information retrieval that tries to overcome these problems by mapping documents as well as terms to a representation in the so-called *latent semantic space*. LSA usually takes the (high dimensional) vector space representation of documents based on term frequencies as a starting point, hence applies a dimension reducing linear projection. The specific form of this mapping is determined by a given document collection and is based on a *singular value decomposition* (in short, SVD) of the corresponding term-document matrix [2].

The general claim is that similarities between documents, or between documents and queries, can be more reliably estimated in the reduced latent space representation than in the original representation. The most important outcome is that documents which share frequently co-occurring terms will have a similar representation in the latent space, even if they have no terms in common. Thus, LSA performs as a kind of noise reduction and has the potential benefit to detect synonyms as well as words that refer to the same topic, even if this relationship is quite implicit and not manifestly identified (as in ESA) [2].

The key idea of LSA is to map documents (and, by symmetry, terms) to a vector space of reduced dimensionality, i.e., the latent semantic space. This mapping is given by decomposing the term-document matrix, say N, through SVD method, in the canonical factorization $N = U\Sigma^t V$, where U and V are orthogonal matrices, i.e., $U^tU = V^tV = I$, while the diagonal matrix Σ contains the singular values of N. In latent semantic indexing, the original vector space representation of documents is replaced by a representation in the low-dimensional latent space, and the similarity is computed based on that representation [2].

In passing, we briefly say as well that, the above approach of semantic decomposition postulates that the semantic primes form a natural semantic metalanguage (in short, NSM) and have a predefined meaning. This meaning is then used to reason the meaning of more complex concepts, which are decomposed. This decomposition will then be used to classify concepts within their meaning and create equivalence classes.

Explicit semantic analysis (ESA), instead, tries to indexing or classify a given document text with respect to a set of explicitly given external categories, already codified a priori. It is just in this sense that ESA is explicit compared to approaches which aim at representing texts with respect to latent (or implicit) topics or concepts, as done in LSA. ESA takes, as input, a document text and maps it to an high-dimensional real-valued vector space (said to be *concept space*). This vector space is spanned by a preassigned database of categories explained in a certain language, such that each dimension corresponds to a category. ESA is explicit in the

*Email: giuseppe.iurato@community.unipa.it

sense that the semantic concept space corresponds exactly to the space of categories taken as a semantic codex providing explicit (or manifest) interpretations.

LSA is a fully automatic mathematical-statistical technique for extracting and inferring relationships of expected contextual usage of words in the various passages of a discourse. It is not a traditional NLP or artificial intelligence program; it uses no humanly already constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies, or other, and takes as its input only raw text parsed into words defined as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs [3].

In structural linguistics, the rules of language are implicitly or tacitly possessed by fluent speakers as a competence, without a conscious, explicit, awareness of how we process, generate and understand language, as a performance, in *normal language processing* (in short, NLP). LSA and ESA are powerful statistical tools which allow us to identify such features. In particular, these statistical analysis methods help us to estimate, in a given document, the replaceability, due to their semantic equivalence, of words in larger text segments. They often use the vector space model, whose dimensionality is related with the number of different terms present in the text [1], [3].

Therefore, we can define a semantic dimensionality of the related vector space as the number of distinct topics represented in it, whose number is much lower than the number of terms. In a given collection, even if both are remarkable features, untapped synonymy is yet more important than unnoticed polysemy. Moreover, the dialectic occurrence vs. co-occurrences of words in a text collection, can tell us what the documents are about and distinguish different senses of polysemous words [1], [3].

Once we have factorized the original document-term matrix using SVD, we can then find a much dimensional smaller matrix that approximates it, roughly, by deleting coefficients from the diagonal matrix, starting with the smallest. These techniques, so to speak, "squeeze down" the matrix to lower rank (typically, around 100-300) by bringing together terms that have similar co-occurrence patterns [3].

The vectors, in this reduced dimensionality space, aren't directly identifiable as any lexical or semantic component, but they are "latently" semantic in that relationships between vectors in this lower dimensional space reflect semantic associations which are not manifestly interpretable on the basis of a given semantic codex of interpretation, as in ESA. Vice versa, the inverse direction to this process of dimensional reduction[1], i.e., the increasing of rank dimensionality of the document-term matrix, is typical of ESA [1], [3].

Reducing the dimensionality of the document-term matrix means, therefore, we are discarding some of the descriptors applied to each document in the collection, which might suggest that retrieval precision would suffer. But, really, we're not just discarding terms, because we are replacing sets of co-occurring (e.g., associated) terms with "superterms", or "topics", that represent predominant (or pregnant) meanings as a kind of average of all the terms that tend to occur in the same contexts. So, we can compute document similarity based on the usual inner product/cosines (similarity) trick in this latent semantic (or concept) space [1], [3].

So, LSA has been shown to be a practical technique for estimating the substitutability, for semantic equivalence, of words in larger text segments. In addition, some of its proponents (like, Susan Dumais) view it as a model of the computational processes and representations underlying substantial portions of how knowledge is acquired and used. And while it is highly unlikely that the conscious human brain uses the same mathematical algorithms as LSA/SVD, it is almost certain that instead the unconscious human brain uses as much analytic power to transform temporally localized experiences into synthesized knowledge just through LSA/SVD [1], [3].

On the other hand, SVD is a well-known algebraic technique which reduces to irreducible

[1] An aspect also characterizing the so-called *semantic differential*.

minimal orthogonal dimensions a given formal problem of linear algebra, so, in the case of document investigation, the reduction to latent semantic space stands out, via LSA, those intrinsic semantic relationships which would not otherwise be identifiable via ESA, hence highlighting those latent, or implicit, links underlying text. Therefore, the emersion of a semantic meaning seems to arise while dimensions of linear spaces, moulding semantic universes within LSA/ESA frameworks, increase. So, the more the linear dimensionality of semantic spaces reduces, the more the text-document latent meaning relations show up.

Therefore, with a deep and well-performed LSA of a text or document, we may highlight those relationships of semantic association which appear to be implicit, or tacit, from an ESA standpoint. This because such latent relatedness is nothing but the *superficial* result of unconscious mechanisms – as just indicated above – rather than cognitive ones, mainly because of the absence of any preassigned codex to which making reference as in ESA. Indeed, while it seems highly doubtful that the human conscious brain uses the same mathematical algorithms as LSA/SVD, it seems almost certain that the brain uses as much analytic power as LSA to transform its temporally local experiences into global knowledge [3].

We put forward the hypothesis that all this takes place in the unconscious realm, at least superficially. To be precise, such latent relationships and correlations emerging from LSA, and that we might suppose to belong to *priming semantics*, are the outcomes of a *displacement* and *condensation* processes, the two main mechanisms of human unconscious, while LSA method might even to be considered as a formal pattern moulding the well-known *free association method* of Freudian psychoanalysis. What we may detect as emerging from unconscious realm is what is placed in its superficial level, in its neighborhoods, because the deepness of unconscious is humanly incognizable, ineffable; only its superficial productions may be picked up by us [4], Vol. II, pp. 549-550.

In conclusion, a deeper investigation between LSA and Freudian psychoanalysis foundations, would deserve to be pursued further.

References

[1] R.J. Glushko, "Module 24. Dimensionality Reduction & Latent Semantic Analysis", Course INFO 202 (19 November 2008), School of Information, University of California, Berkeley, CA, 2008.

[2] T. Hoffmann, "Probabilistic Latent Semantic Indexing", in: K.B. Laskey, H. Prade (Eds.), *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence – UAI'99*, Morgan Kaufmann Publishers, Inc., San Francisco, CA, 1999, pp. 289-296.

[3] T.K. Landauer, P.W. Foltz, D. Laham, "Introduction to Latent Semantic Analysis", *Discourse Processes*, 25 (1998) pp. 259-284.

[4] M. Recalcati, *Jacques Lacan*, 2 vols., Raffaello Cortina Editore, Milano, 2012-16.