

Semantic Spaces: Measuring the Distance between Different Subspaces

Guido Zuccon¹ Leif Azzopardi¹ C. J. van Rijsbergen¹

¹Department of Computing Science, University of Glasgow, Scotland (UK)



Third Quantum Interaction Symposium 2009

Background

Motivation

Comparing Semantic Spaces

Subspace distance for Semantic Spaces

Conclusion and Future Work

Background: Semantic Spaces and QT

- ▶ Semantic Spaces map words into a high dimensional vector space
- ▶ They are build by computing lexical co-occurences between words appearing in the same context (window of text)
- ▶ In [Bruza and Woods, 2008] Semantic Spaces (HAL) are modeled by density operators: the collapse of states represents the collapse into a specific meaning of a word

Background: Semantic Spaces and QT

- ▶ The density operator associated to a Semantic Space contains mainly two information: the states of the system and the probability distribution used to prepare the system
- ▶ In QM, the need to distinguish between different preparations of systems has emerged
- ▶ This is not a trivial problem: because of the statistical error introduced when measuring frequencies of outcomes, it is difficult to distinguish between preparations of the same system that slightly differ [Wootters, 1981]

Background: Semantic Spaces and QT

- ▶ [Wootters, 1981] suggested to use statistical fluctuations to determine the distance between states of different preparation of quantum systems.
- ▶ Results obtained by the statistical approach are the same obtained by measuring the angles between rays in a Hilbert space (associated to the quantum states).
- ▶ This measure is at the basis of the distance we propose in our study.

Motivation of this study (from an experimental viewpoint)

Is it possible to measure the distance between Semantic Spaces (SS) (represented by means of density operators)?

- ▶ the outcome of the distance would tell if two SS represent the same “system” or not.

Motivation of this study (from an information retrieval viewpoint)

Why comparing SS? Why a distance between them?

- ▶ comparing the SS of documents provides a measurement of semantic similarity between documents
- ▶ if the cluster hypothesis holds for the SS distance, then documents related to the same concepts will be close to each other w.r.t. the SS distance;

A distance between word representations in SS can be calculated for example by means of Minkowski distance.

However, no global distance has been investigated to compare SS.

Comparing Semantic Spaces – Pairwise

How do we compare two Semantic Spaces?

- ▶ Pairwise comparison: consider how each word is represented, e.g.:
 1. a word w is represented by
$$w_1 = [0.1678, 0.4196, 0, 0.5874, 0.6713]$$
 in H_1 and by
$$w_2 = [0, 0.8660, 0.2887, 0.2887, 0.2887]$$
 in H_2
 2. we compare the two word representations with a local distance, for example Euclidean distance:

$$\text{dist}(w_1, w_2) = \sqrt{\sum_{i=1}^n (w_1(i) - w_2(i))^2} = 0.7393$$

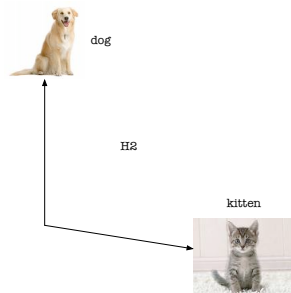
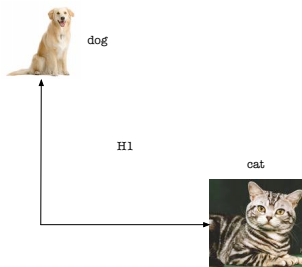
3. this procedure is applied to each pair of the same word in the two spaces: we are comparing the word representations
- ▶ Global comparison: consider the geometry of the subspaces

Comparing Semantic Spaces – Global

How do we compare two Semantic Spaces?

- ▶ Pairwise comparison: consider how each word is represented
- ▶ Global comparison: consider the geometry of the subspaces
 1. we don't compare the word representation word-by-word
 2. we compare an entire subspace representation against another
 3. takes into account the whole geometry of the subspace built by a set of documents, not the geometry of a word
 4. we take into account the concepts expressed in a set of documents by getting the basis of the SS and then measure the distance between two SS, representing the distance between the concepts/contexts expressed in two sets of documents.

Pairwise VS Global Comparison



Properties of a SS distance

Desired properties of a SS distance:

- ▶ invariant to the choice of the basis
- ▶ take into account difference in the subspaces dimensions

The subspace distance

A possible candidate is the Subspace distance:

$$\text{distance}(H_{d_1}, H_{d_2}) = \sqrt{\max(m, n) - \text{tr}(Q_2^T Q_1 Q_1^T Q_2)}$$

where:

- ▶ $\text{rank}(H_{d_1}) = m$ and $\text{rank}(H_{d_2}) = n$
- ▶ Q_1 orthonormal (orthogonal and normalized) basis for H_{d_1}
- ▶ Q_2 orthonormal basis for H_{d_2}

The subspace distance – properties

- ▶ the distance is invariant to the choice of the orthonormal basis (hint: use Parseval's theorem for demonstration)
- ▶ it is symmetric
- ▶ non-negative ($distance(H_{d_1}, H_{d_2}) \geq 0$) and $distance(H_{d_1}, H_{d_2}) = 0 \iff H_{d_1} = H_{d_2}$
- ▶ upper boundedness: $distance(H_{d_1}, H_{d_2}) \leq \sqrt{(\max(m, n))}$, and $distance(H_{d_1}, H_{d_2}) = \sqrt{(\max(m, n))} \iff H_{d_1} \perp H_{d_2}$
- ▶ the triangularity inequality holds

The subspace distance

- ▶ The distance has been originally proposed by [Wang et al., 2006] for face recognition
- ▶ has a strong relationship with the chordal distance
$$d_c(S_a, S_b) = \sqrt{m - \text{tr}(P_a P_b)}$$

Experimentation (1/2)

Aim of the experiments: empirically demonstrate that related documents are at closer subspace distance between each other than not related ones.

Settings of the experiment:

- ▶ baseline: Euclidean distance between SS representations of words
- ▶ we use a standard IR collection (WSJ) as source of documents used to generate SS: $\geq 170k$ newswire articles, $\geq 226k$ unique words
- ▶ we randomly selected one Topic from the TREC1 sets.
- ▶ create two subsets of documents: R contains documents judged relevant to the topic, vice versa N contains documents judged irrelevant.

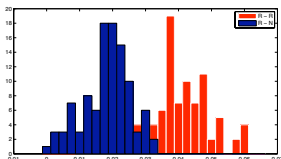
Experimentation (2/2)

Procedure:

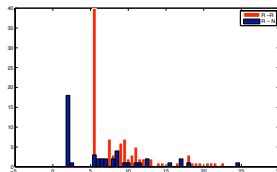
- ▶ compute the SS representation for all the documents in R and N (two methods: HAL or HAL traces)
- ▶ compute the semantic subspace and the euclidean distance between each pair of spaces taken from R only, from N only, and one from R while the other is taken from N.

Results (1/2)

Figure: Frequencies distribution of pairwise subspace distances (a) and Euclidean distances (b) between subspaces belonging to the set of relevant documents (R) and the non relevant (N).



(a) Distribution of frequencies for the subspace distance.



(b) Distribution of frequencies for the Euclidean distance.

Thus:

- ▶ Spaces $\in R$ are on average at closer subspace distance to each other than to spaces $\in N$.

Results (2/2)

Table: Average distance between sets of relevant documents (R) and not relevant documents (N) obtained by the *Subspace* distance (values reported refer to $1 - d_s(S_a, S_b) / \sqrt{\max(p, r)}$) and the *Euclidean* distance.

	R		N	
	Subspace	Euclidean	Subspace	Euclidean
R	0.0376 ± 0.0116	9.3910 ± 4.6994	0.0182 ± 0.0072	6.7059 ± 5.5936
N	0.0182 ± 0.0072	6.7059 ± 5.5936	0.0386 ± 0.0093	3.3816 ± 2.0667

Conclusion

- ▶ In this investigation we have proposed a measure for comparing Semantic Spaces, providing empirical results.
- ▶ The proposed measure is a possible Semantic Space counterpart of the statistical distance between preparations of quantum system in QT.
- ▶ Drawback: high computational time requested to compute the subspace distance over standard IR collections, making the distance unsuitable for online classification of documents.

Future Work

Future work will consider:

- ▶ Investigate alternative formulation of the measure or exploit the geometry underneath the distance to provide an efficient implementation in terms of computational time.
- ▶ Experimentations directed to apply the measure in a number of retrieval applications in order to determine its effectiveness.

Bibliography I

- [Bruza and Woods, 2008] Bruza, P. D. and Woods, J. (2008).
Quantum Collapse in Semantic Space: Interpreting Natural Language
Argumentation.
In 2nd QI Symposium, pages 141–147.
- [Wang et al., 2006] Wang, L., Wang, X., and Feng, J. (2006).
Subspace Distance Analysis with Application to Adaptive Bayesian Algorithm for
Face Recognition.
Pattern Recognition, 39(3):456–464.
- [Wootters, 1981] Wootters, W. K. (1981).
Statistical Distance and Hilbert Space.
Phys. Rev. D, 23(2):357–362.