Investigating Subspace Distances in Semantic Spaces

G. Zuccon[1], L. Azzopardi[1], and E. Gasco[2]
[1] School of Computing Science, University of Glasgow, Scotland (UK)
[2] Zirak, Mondoví, Italy
{guido, leif}@dcs.gla.ac.uk, enrico@zirak.it

Abstract Semantic space models of word meaning derived from co-occurrence statistics within a corpus of documents, such as the Hyperspace Analogous to Language (HAL) model, have been proposed in the past. While word similarity can be computed using these models, it is not clear how semantic spaces derived from different sets of documents can be compared. In this paper, we focus on this problem, and we revisit the proposal of using semantic subspace distance measurements [1]. In particular, we outline the research questions that still need to be addressed to investigate and validate these distance measures. Then, we describe our plans for future research.

1 Introduction

Traditional information retrieval (IR) paradigms use statistics of word occurrence to match documents with users queries. Although a number of IR models have been proposed, they usually focus on key word matching [2]. However, there are many occasions when the queries do not contain the exact same terms which appear in relevant documents. Consequently, models that only consider word occurrence statistics will fail to rank such documents highly, if at all [3]. It has been argued that some of the limits of IR systems can be overcome by understanding the "meaning" of the user query and of the indexed documents [4]. Mathematical and computational models of (word) meaning and semantics have been proposed in the past. Among those, semantic space models have attracted much interest and shown to be successful in particular tasks, e.g. information inference for query expansion [5]. Common to semantic space techniques is the mapping of words into a high dimensional vector space [6], generally obtained by computing[3] lexical co–occurrences between words appearing in the same context, e.g. Hyperspace Analogue to Language (HAL) [7] and probabilistic HAL (pHAL) [8] approaches. Other models however can be used that do not derive semantic evidences directly from word co-occurrences, e.g. Latent Semantic Analysis [9]. However, a problem common to all types of semantic spaces is how to measure the distance between subspaces. Previous work has been limited to measuring distances between vector representations of words within the same semantic space. For example, the Minkowski distance, or alternatively one of its specialisation, such as the Euclidean distance, can be used to compare individual word

[3] Vectors are assigned to words so as to represent the co–occurrences between a word and others.

vectors (see [7]). Following this approach in order to measure distance between semantic spaces, however, would led to imprecise measurements. This is because a concept can be expressed by different words in different semantic spaces: thus their vector representations are not going to be considered. In [1], a measure has been developed, inspired by Quantum Physics, to overcome this problem. The approach compares whole subspaces (generated from semantic spaces), instead of individual vectors, i.e. documents are represented with subspaces instead of word-vectors. This provides a novel way in which the distance measurements can be performed, as all the vectors that form a basis of the semantic spaces contribute in building up the distance measurement.

In this paper we build upon the proposal suggested in [1], and outline the plan of a joint collaboration between the University of Glasgow and Zirak[4]. This industry funded project aims to explore the subspace distance measure and empirically investigate its utility and application within IR.

2 Measuring the Distance between Subspaces

The semantic subspace distance proposed in [1] is inspired by the chordal distance, a monotonic function of the inner product. The idea underlying this measure is that all the basis vectors that describe the subspaces should concur to build up the distance between the subspaces themselves. This is in contrast with the use of pairwise distance measurements between representations of the same word in different subspaces. The definition of the semantic subspace distance can be given either in relation to the inner product between vectors belonging to the bases (i.e. $\mathbf{u_i}$, $\mathbf{v_i}$) of the compared subspaces, or depending upon the trace (indicated by the function $tr(.)$) of the projectors (i.e. P_a, P_b) describing the subspaces:

$$d_s(S_a, S_b) = \sqrt{\max(p, r) - \sum_{i=1}^{p}\sum_{j=1}^{r}(\mathbf{u_i}^T\mathbf{v_j})^2} \quad (1)$$

$$= \sqrt{\max(p, r) - tr(P_a P_b)} \quad (2)$$

where p and r are the dimensions[5] of the subspaces S_a and S_b, respectively. Geometrically, this "loosely"[6] corresponds to take into account the projection of a subspace into another subspace (see eq 2), rather than the intersection between two subspaces. The semantic subspace distance can be employed to compute the distance between semantic subspaces (as, for example, the ones generated by HAL) aiming to obtain a more precise measurement of separation than using a naïve distance based upon the comparison between single word representations, e.g. the Minkowski distance. We refer the interested reader to [1] for the derivation and a discussion of the properties of this distance.

In [1], a pilot study was performed which showed that the measure grouped relevant documents at a closer distance than irrelevant documents. Thus, the

[4] `http://www.zirak.it/`

[5] The dimension of a subspace is defined as the number of vectors that form the basis of the subspace; recall that a basis must both span the subspace and be linearly independent.

[6] In the sense that eq 2 considers also the dimensionality of the subspaces and applies a trace operator to the projection of the subspaces.

distance might be better at discriminating between relevant and non relevant documents than other similarity measures. While these initial results are encouraging, they warrant further investigation, in particular by testing the distance measure on a number of heterogeneous document collections. Furthermore, neither computational efficiency problems nor the suitability for real IR applications have been investigated and discussed. In the next section we outline a number of open questions that have yet to been investigated and our plan to address those problematics.

3 Directions of Future Research

In the following we detail the research questions and the plan of investigation of a joint project between the University of Glasgow and Zirak, that aims to develop further knowledge about the semantic subspace distance. The final outcome of the project is two-fold. On one hand we aim to achieve a deeper understanding about the geometry of semantic spaces and about the semantic subspace measure. On the other hand, we aim to release an open source package for creating semantic subspaces and measuring subspaces distances, so to provide the IR community with a tool for using the distance in real retrieval applications, such as information filtering. Our research questions are:

1. which semantic models can be used to obtain a semantic subspace?
2. how can a basis for a semantic subspace be found? Is there a technique that provides a (qualitative) "better" basis than alternative techniques?
3. how can the distance be tested?
4. in which IR scenarios can the measure be applied?

Semantic models for semantic (sub)spaces. Previous studies have used HAL for deriving semantic spaces from sets of documents. Other techniques might be used to derive alternative representations of documents in terms of word co-occurrences and semantics [8,10]. However, we have chosen to focus our implementation and investigation on HAL and probabilistic HAL [8].

Finding a basis for semantic (sub)spaces. A common linear algebra procedure for finding the basis of a subspace is the application of the Gram-Schimdt algorithm (which produces an *orthonormal* basis). However, alternative algorithms are suitable for this problem. Furthermore, given the large dimensionality of semantic subspaces and the success of techniques for dimensionality reduction within IR, algorithms for dimensionality reduction can be employed to extract an approximate (reduced) basis. For example, Singular Value Decomposition yields to matrices that contain a set of orthonormal basis vectors (the eigenvectors) for the subspace, as well as a matrix containing the eigenvalues associated to the basis vectors with respect to the given subspace. In preliminary studies, the QR decomposition (which directly implements the Gram-Schmidt algorithm) has been employed to obtain subspaces bases. If dimensionality reduction is not applied, there are no quantitative differences with respect to the semantic subspace distance in the choice of any of the previous techniques. Therefore, we not only plan to investigate one of these techniques, but also a particular matrix decomposition procedure (the Non-negative Matrix Factorisation – NMF), that indeed yields qualitative as well as quantitative differences when

computing the semantic subspace distance. In fact, NMF generates bases that are *not orthonormal* (nor orthogonal), but so that the elements of the bases vectors are positive (or zero). Qualitatively, such bases might have interpretative advantages with respect to bases that also contain negative numbers[7].

Testing the semantic subspace distance. It is our plan to extend the pilot study of the semantic subspace distance reported in [1], by considering a greater and heterogeneous set of collections. Furthermore, we are interested in experimenting with semantic spaces created from texts and news feeds written in languages other than English, in particular Italian. However, this might require the creation of a suitable test collections that contains topics and associated relevance judgments. Indeed, experiments performed on these collections will enable us to obtain an indication of the goodness of the distance measure.

IR tasks for the semantic subspace distance. Preliminary results suggest that the semantic subspace distance is able to discriminate between relevant and irrelevant documents. If this finding is confirmed during the planned investigation, then the distance might be successfully used in the tasks of information filtering and for relevance feedback, where a set of relevant documents can be given as input to the system. A further area of application is that of document classification and clustering, where the semantic subspace distance can be used for determining how the similarity of two documents is calculated.

Acknowledgments This project is funded and supported by Zirak under grant agreement no. 55500/1 and by the EPSRC Renaissance (EP/F014384/1) project.

References

1. Zuccon, G., Azzopardi, L., van Rijsbergen, C.: Semantic spaces: Measuring the distance between different subspaces. In: QI. Volume 5494. (2009) 225–236
2. Spark-Jones, K., Robertson, S., Hiemstra, D., Zaragoza, H.: Language modelling and relevance. In: Language Modeling for Information Retrieval. Kluwer Academic Publishers (2003)
3. Belkin, N.J., Oddy, R.N., Brooks, H.M.: Ask for information retrieval: Part I. background and theory. Journal of Documentation **38**(2) (1982) 61–71
4. Furnas, G.W., Deerwester, S., Dumais, S.T., Landauer, T.K., Harshman, R.A., Streeter, L.A., Lochbaum, K.E.: Information retrieval using a singular value decomposition model of latent semantic structure. In: SIGIR '88. (1988) 465–480
5. Song, D., Bruza, P.D.: Discovering Information Flow Using High Dimensional Conceptual Space. In: SIGIR '01. (2001) 327–333
6. Osgood, C., Suci, G., Tannenbaum, P., Date, P.: The Measurement of Meaning. University of Illinois Press (US) (1957)
7. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments & Computers (1996)
8. Azzopardi, L., Girolami, M., Crowe, M.: Probabilistic hyperspace analogue to language. In: SIGIR '05. (2005) 575–576
9. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by Latent Semantic Analysis. JASIS **41**(6) (1990) 391–407
10. Jurgens, D., Stevens, K.: The s-space package: an open source package for word space models. In: ACL 2010. (2010) 30–35

[7] If the value of a feature is "negative", this is difficult to interpret, as it does not necessarily imply the absence, or negation, of that feature.