Dimensionality Reduction for Text using Domain Knowledge

Yi Mao and **Krishnakumar Balasubramanian** and **Guy Lebanon**
Georgia Institute of Technology

Abstract

Text documents are complex high dimensional objects. To effectively visualize such data it is important to reduce its dimensionality and visualize the low dimensional embedding as a 2-D or 3-D scatter plot. In this paper we explore dimensionality reduction methods that draw upon domain knowledge in order to achieve a better low dimensional embedding and visualization of documents. We consider the use of geometries specified manually by an expert, geometries derived automatically from corpus statistics, and geometries computed from linguistic resources.

1 Introduction

Visual document analysis systems such as IN-SPIRE have demonstrated their applicability in managing large text corpora, identifying topics within a document and quickly identifying a set of relevant documents by visual exploration. The success of such systems depends on several factors with the most important one being the quality of the dimensionality reduction. This is obvious as visual exploration can be made possible only when the dimensionality reduction preserves the structure of the original space, i.e., documents that convey similar topics are mapped to nearby regions in the low dimensional 2D or 3D space.

Standard dimensionality reduction methods such as principal component analysis (PCA), locally linear embedding (LLE) (Roweis and Saul, 2000), or t-distributed stochastic neighbor embedding (t-SNE) (van der Maaten and Hinton, 2008) take as input a set of feature vectors such as bag of words. An obvious drawback is that such methods ignore the textual nature of documents and instead consider the vocabulary words $v_1, \ldots, v_n$ as abstract orthogonal dimensions.

In this paper we introduce a framework for incorporating domain knowledge into dimensionality reduction for text documents. Our technique does not require any labeled data, therefore is completely unsupervised. In addition, it applies to a wide variety of domain knowledge.

We focus on the following type of non-Euclidean geometry where the distance between document x and y is defined as

$$d_T(x, y) = \sqrt{(x-y)^\top T(x-y)}. \qquad (1)$$

Here $T \in \mathbb{R}^{n\times n}$ is a symmetric positive semidefinite matrix, and we assume that documents x, y are represented as term-frequency (tf) column vectors. Since T can always be written as $H^\top H$ for some matrix $H \in \mathbb{R}^{n\times n}$, an equivalent but sometimes more intuitive interpretation of (1) is to compose the mapping $x \mapsto Hx$ with the Euclidean geometry

$$d_T(x, y) = d_I(Hx, Hy) = \|Hx - Hy\|_2. \qquad (2)$$

We can view T as encoding the semantic similarity between pairs of words and H as smoothing the tf vector by mapping observed words to related but unobserved words. Therefore, the geometry realized by (1) or (2) may be used to derive novel dimensionality reduction methods that are customized to text in general and to specific text domains in particular. The main challenge is to obtain the matrices H or T that describe the relationship among vocabulary words appropriately.

We consider three general ways of obtaining H or T using domain knowledge. The first corresponds to manually specifying H or T based on the semantic relationship among words (determined by domain expert). The second corresponds to constructing H or T by analyzing relationships between different words using corpus statistics. The third is based on knowledge obtained from linguistic resources. Whether to specify H directly or indirectly by specifying $T =$

Coling 2010: Poster Volume, pages 801–809,
Beijing, August 2010

$H^\top H$ depends on the knowledge type and is discussed in detail in Section 4.

We investigate the performance of the proposed dimensionality reduction methods for three text domains: sentiment visualization for movie reviews, topic visualization for newsgroup discussion articles, and visual exploration of ACL papers. In each of these domains we evaluate the dimensionality reduction using several different quantitative measures. All the techniques mentioned in this paper are unsupervised, making use of labels only for evaluation purposes.

Our take home message is that all three approaches mentioned above improves dimensionality reduction for text upon standard embedding ($H = I$). Furthermore, geometries obtained from corpus statistics are superior to manually constructed geometries and to geometries derived from standard linguistic resources such as WordNet. Combining heterogenous types of knowledge provides the best results.

2 Related Work

Despite having a long history, dimensionality reduction is still an active research area. Broadly speaking, dimensionality reduction methods may be classified as projective or manifold based (Burges, 2009). The first projects data onto a linear subspace (e.g., PCA and canonical correlation analysis) while the second traces a low dimensional nonlinear manifold on which data lies (e.g., multidimensional scaling, isomap, Laplacian eigenmaps, LLE and t-SNE). The use of dimensionality reduction for text documents is surveyed by Thomas and Cook (2005) who also describe current homeland security applications.

Dimensionality reduction is closely related to metric learning. Xing et al. (2003) is one of the earliest papers that focus on learning metrics of the form (1). In particular they try to learn matrix T in an supervised way by expressing relationships between pairs of samples. A representative paper on unsupervised metric learning for text documents is Lebanon (2006) which learns a metric on the simplex based on the geometric volume of the data.

We focus in this paper on visualizing a corpus of text documents using a 2-D scatter plot. While this is perhaps the most popular and practical text visualization technique, other methods such as Spoerri (1993), Hearst (1997), Havre et al. (2002), Paley (2002), Blei et al. (2003), Mao et al. (2007) exist. Techniques developed in this paper may be ported to enhance these alternative visualization methods as well.

3 Non-Euclidean Geometries

Dimensionality reduction methods often assume, either explicitly or implicitly, Euclidean geometry. For example, PCA minimizes the reconstruction error for a family of Euclidean projections. LLE uses the Euclidean geometry as a local metric. t-SNE is based on a neighborhood structure, determined again by the Euclidean geometry. The generic nature of the Euclidean geometry makes it somewhat unsuitable for visualizing text documents as the relationship between words conflicts with Euclidean orthogonality. We consider in this paper several alternative geometries of the form (1) or (2) which are more suited for text and compare their effectiveness in visualizing documents.

As mentioned in Section 1, H smooths the tf vector x by mapping the observed words into observed and non-observed (but related) words. In case H is nonnegative, it can be further decomposed into a product of a non-negative column normalized matrix $R \in \mathbb{R}^{n\times n}$ and a non-negative diagonal matrix $D \in \mathbb{R}^{n\times n}$. The decomposition $H = RD$ shows that H has two key roles. It smooths related vocabulary words (realized by R) and it emphasizes some words over others (realized by D). Setting R_{ij} to a high value if w_i, w_j are similar and 0 if they are unrelated maps an observed word to a probability vector over related words in the vocabulary. The value D_{ii} captures the importance of v_i and therefore should be higher for important content words than for less important words or stop-words[1].

It is instructive to examine the matrices R and D in the case where the vocabulary words cluster in some meaningful way. Figure 1 gives an example where vocabulary words form two clusters. The matrix R may become block-diagonal with non-zero elements occupying diagonal blocks representing within-cluster word

[1] The nonnegativity assumption of H is useful when constructing H by domain experts such as the method A in Section 4. In general, H needs not to be nonnegative for dimensionality reduction as in (2).

$$\begin{pmatrix} 0.8 & 0.1 & 0.1 & 0 & 0 \\ 0.1 & 0.8 & 0.1 & 0 & 0 \\ 0.1 & 0.1 & 0.8 & 0 & 0 \\ 0 & 0 & 0 & 0.9 & 0.1 \\ 0 & 0 & 0 & 0.1 & 0.9 \end{pmatrix} \begin{pmatrix} 5 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

Figure 1: An example of a decomposition $H = RD$ in the case of two word clusters $\{v_1, v_2, v_3\}$, $\{v_4, v_5\}$. The block diagonal elements in R represent the fact that words are mostly mapped to themselves, but sometimes are mapped to other words in the same cluster. The diagonal matrix indicates that the first cluster is more important than the second cluster for the purposes of dimensionality reduction.

blending, i.e., words within each cluster are interchangeable to some degree. The diagonal matrix D represents the importance of different clusters. The word clusters are formed with respect to the visualization task at hand. For example, in the case of visualizing the sentiment content of reviews we may have word clusters labeled as "positive sentiment words", "negative sentiment words" and "objective words".

In general, the matrices R, D may be defined based on the language or may be specific to document domain and visualization purpose. It is reasonable to expect that the words emphasized for visualizing topics in news stories might be different than the words emphasized for visualizing writing styles or sentiment content.

Applying the geometry (1) or (2) to dimensionality reduction is easily accomplished by first mapping document tf vectors $x \mapsto Hx$ and proceeding with standard dimensionality reduction techniques such as PCA or t-SNE. The resulting dimensionality reduction is Euclidean in the transformed space but non-Euclidean in the original space. In many cases, the vocabulary contains tens of thousands of words or more making the specification of T or H a complicated and error prone task. We describe in the next section several techniques for specifying these matrices in practice.

4 Domain Knowledge

Method A: Manual Specification

In this method, a domain expert manually specifies $H = RD$ by specifying (R, D) based on the perceived relationship among the vocabulary words. More specifically, the user first constructs a hierarchical word clustering that may depend on the current text domain, and then specifies the matrices (R, D) based on the clustering.

Denoting the clusters by $C_1, \ldots, C_r$ (a partition of $\{v_1, \ldots, v_n\}$), R is set to

$$R_{ij} \propto \begin{cases} \rho_a, & i = j, v_i \in C_a \\ \rho_{ab}, & i \neq j, v_i \in C_a, v_j \in C_b \end{cases}.$$

The values $\rho_{ab}, a \neq b$ capture the semantic similarity between two clusters and the value ρ_{aa} captures the similarity of two different words within the cluster a. These values may be set manually by domain expert or automatically computed based on the clustering hierarchy (for example ρ_{ab} can be the inverse of the minimal number of tree edges traversed in moving from a to b). To maintain a probabilistic interpretation, the matrix R should be normalized so that its columns sum to 1. The diagonal matrix D is specified by setting the values

$$D_{ii} = d_a, \quad v_i \in C_a$$

according to the importance of word cluster C_a to the current visualization task.

We emphasize that as with the rest of the methods in this paper, the manual specification is done without access to labeled data. Since manual clustering assumes some form of human intervention, it is reasonable to also consider cases where the user specifies H or T in an interactive manner. For example, the expert specifies an initial clustering of words and values for (R, D), views the resulting embeddings and adjusts the selection interactively until reaching a satisfactory embedding.

Method B: Contextual Diffusion

An alternative to manually specifying $T = DR^\top RD$ is to construct it based on similarity between the contextual distributions of the vocabulary words. The contextual distribution of word v is defined as

$$q_v(w) = p(w \text{ appears in } x | v \text{ appears in } x) \quad (3)$$

where x is a randomly drawn document. In other words q_v is the distribution governing the words appearing in the context of word v.

A natural similarity measure between distributions is the Fisher diffusion kernel proposed by Lafferty and Lebanon (2005). Applied to contextual distributions as in Dillon et al. (2007) we arrive at the following similarity matrix

$$T(u,v) = \exp\left(-c \arccos^2\left(\sum_w \sqrt{q_u(w) q_v(w)}\right)\right).$$

where $c > 0$. Intuitively, the word u will be diffused into v depending on the geometric diffusion between the distributions of likely contexts.

We use the following formula to estimate the contextual distribution from a corpus

$$\begin{aligned} q_v(w) &= \sum_{x'} p(w, x'|v) = \sum_{x'} p(w|x', v) p(x'|v) \\ &= \sum_{x'} \text{tf}(w, x') \frac{\text{tf}(v, x')}{\sum_{x''} \text{tf}(v, x'')} \qquad (4) \\ &= \left(\frac{1}{\sum_{x'} \text{tf}(v, x')}\right)\left(\sum_{x'} \text{tf}(w, x')\text{tf}(v, x')\right) \end{aligned}$$

where tf(w, x) is the number of times word w appears in document x divided by the length of the document x. The contextual distribution q_v or diffusion matrix T above may be computed in an unsupervised manner without labels.

Method C: Web n-Grams

In method B the contextual distribution is computed using a large external corpus that is similar to the text being analyzed. An alternative that is especially useful when such a corpus is not easily available is to use generic resources to estimate the contextual distribution (3)-(4). One option is to use the publicly available Google n-gram dataset (Brants and Franz, 2006) to estimate T. More specifically, we compute the contextual distribution by considering the proportion of times two words appear together within the n-grams e.g., for $n = 2$ we have

$$q_v(w) = \frac{\text{\# of bigrams containing both } w \text{ and } v}{\text{\# of bigrams containing } v}.$$

Method D: Word-Net

In the last method, we consider using Word-Net, a standard linguistic resource, to specify T. This is similar to manual specification (method A) in that it builds upon experts' knowledge rather than corpus statistics. In contrast to method A, however, Word-Net is a carefully built resource containing more accurate and comprehensive linguistic information such as synonyms, hyponyms and holonyms. On the other hand, its generality puts it at a disadvantage as method A may be adapted to a specific text domain.

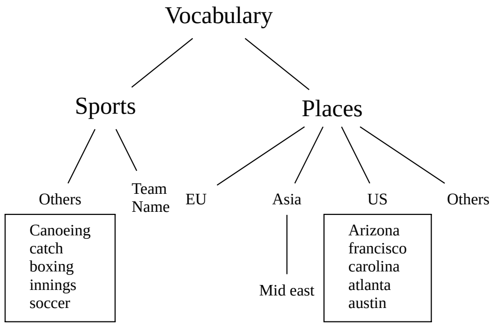


Figure 2: Manually specified hierarchical word clustering for the 20 newsgroup domain. The words in the frames are examples of words belonging to several bottom level clusters.

We follow Budanitsky and Hirst (2001) who compared five similarity measures between words based on Word-Net. In our experiments we use the measure of Jiang and Conrath (1997) (see also Jurafsky and Martin (2008))

$$T(u,v) = \log \frac{p(u)p(v)}{2p(\text{lcs}(u,v))}$$

as it was shown to outperform the others. Above, lcs stands for the lowest common subsumer, i.e., the lowest node in the hierarchy that subsumes (is a hypernym of) both u and v. The quantity $p(u)$ is the probability that a randomly selected word in a corpus is an instance of the synonym set that contains word u.

Combination of Methods

In addition to individual methods we also consider their convex combinations

$$H^* = \sum_i \alpha_i H_i \quad \text{s.t.} \quad \alpha_i \geq 0, \sum_i \alpha_i = 1 \quad (5)$$

where H_i are matrices from methods A-D (obtained implicitly by specifying R and D for method A and T for methods B-D). Doing so allows us to combine heterogeneous types of domain knowledge including experts' knowledge and corpus statistics, leverage their diverse nature and potentially achieve better performance than any of the methods on its own.

5 Experiments

We evaluate the proposed methods by experimenting on two text datasets where domain knowledge is relatively easy to obtain (especially for method A and B). Preprocessing includes lower-casing, stop words removal, stemming, and selecting the most frequent 2000 words for both datasets.

The first is the Cornell sentiment scale dataset of movie reviews from 4 critics (Pang and Lee, 2004). The visualization in this case focuses on the sentiment quantity of either 1 (very bad) or 4 (very good) (Pang et al., 2002). For method A, we use the General Inquirer resource[2] to partition the vocabulary into three clusters conveying positive, negative or neutral sentiment. While visualizing documents from one particular author, the rest of the reviews from other three authors can be used as an estimate of contextual distribution for method B.

The second text dataset is the 20 newsgroups. It consists of newsgroup articles from 20 distinct newsgroups and is meant to demonstrate topic visualization. In this case one of the authors designed a hierarchical clustering of the vocabulary words based on general knowledge of English language (see Figure 2 for a partial clustering hierarchy) without access to labels. The contextual distribution for method B is estimated from the Reuters RCV1 dataset (Lewis et al., 2004) which consists of news articles from Reuters.com in the year 1996 and 1997.

Method C uses Google n-gram which provides a massive scale resource for estimating the contextual distribution. In the case of Word-Net (method D) we used Pedersen's implementation of Jiang and Conrath's similarity measure[3]. Note, for these two methods, the obtained information is not domain specific but rather represents general semantic relationships between words.

In our experiments below we focused on two dimensionality reduction methods: PCA and t-SNE. PCA is a well known classical method while t-SNE (van der Maaten and Hinton, 2008) is a recent dimensionality reduction technique for visualization purposes. The use of t-SNE is motivated by the fact that it was shown to outperform LLE, CCA, MVU, Isomap, and Laplacian eigenmaps when the dimensionality of the data is reduced to two or three.

To measure the dimensionality reduction quality, we visualize the data as a scatter plot with different data groups (topics, sentiments) displayed with different markers and colors. Our quantitative evaluation of the visualization is based on the fact that documents belonging to different groups (topics, sentiments) should be spatially separated in the 2-D space. Specifically, we used the following indices:

(i) The weighted intra-inter criteria is a standard clustering quality index that is invariant to non-singular linear transformations of the embedded data. It equals tr$(S_T^{-1} S_W)$ where S_W is the within-cluster scatter matrix, $S_T = S_W + S_B$ is the total scatter matrix, and S_B is the between-cluster scatter matrix (Duda et al., 2001).

(ii) The Davies Bouldin index is an alternative to (i) that is similarly based on the ratio of within-cluster scatter to between-cluster scatter (Davies and Bouldin, 2000).

(iii) Classification error rate of a k-NN classifier that applies to data groups in the 2-D embedded space. Despite the fact that we are not interested in classification per se (otherwise we would classify in the original high dimensional space), it is an intuitive and interpretable measure of cluster separation.

(iv) An alternative to (iii) is to project the embedded data onto a line which is the direction returned by applying Fisher's linear discriminant analysis to the embedded data. The projected data from each group is fitted to a Gaussian whose separation is used as a proxy for visualization quality. In particular, we summarize the separation of the two Gaussians by measuring the overlap area. While (iii) corresponds to the performance of a k-NN classifier, method (iv) corresponds to the performance of Fisher's LDA classifier.

Labeled data is not used during the dimensionality reduction stage but it is used in each of the above measures for evaluation purposes.

[2] http://www.wjh.harvard.edu/∼inquirer/

[3] http://wn-similarity.sourceforge.net/

Figure 3 displays both qualitative and quantitative evaluation of PCA and t-SNE for the sentiment and newsgroup domains for $H = I$ (left column), manual specification (middle column) and contextual distribution (right column). In general for both domains, methods A and B perform better both qualitatively and quantitatively (indicating by the numbers in the top two rows) than the original dimensionality reduction with method B outperforming method A.

Tables 1-2 compare evaluation measures (i) and (iii) for different types of domain knowledge. Table 1 corresponds to the sentiment domain where we conducted separate experiments for four movie critics. Table 2 corresponds to the newsgroup domain where two tasks were considered. The first involves three newsgroups (comp.sys.mac.hardware vs. rec.sports.hockey vs. talk.politics.mideast) and the second involves four newsgroups (rec.autos vs. rec.motorcycles vs. rec.sports.baseball vs. rec.sports.hockey). It is clear from these two tables that the contextual diffusion, Google n-gram, and Word-Net generally outperform the original $H = I$ matrix. The best method varies from task to task but the contextual diffusion and Google n-gram in general result in good performance.

	PCA (1)	PCA (2)	t-SNE (1)	t-SNE (2)
$H = I$	1.5391	1.4085	1.1649	1.1206
B	1.2570	**1.3036**	1.2182	1.2331
C	**1.2023**	1.3407	**0.7844**	**1.0723**
D	1.4475	1.3352	1.1762	1.1362
	PCA (1)	PCA (2)	t-SNE (1)	t-SNE (2)
$H = I$	0.8461	0.5630	0.9056	0.7281
B	0.7381	**0.6815**	0.9110	0.6724
C	0.8420	0.5898	**0.9323**	0.7359
D	**0.8532**	0.5868	0.9013	**0.7728**

Table 2: Quantitative evaluation of dimensionality reduction for visualization for two tasks in the news article domain. The numbers in the top five rows correspond to measure (i) (lower is better), and the numbers in the bottom five rows correspond to measure (iii) ($k = 5$) (higher is better). We conclude that contextual diffusion (B), Google n-gram (C), and Word-Net (D) tend to outperform the original $H = I$.

We also examined convex combinations

$$\alpha_1 H_A + \alpha_2 H_B + \alpha_3 H_C + \alpha_4 H_D \qquad (6)$$

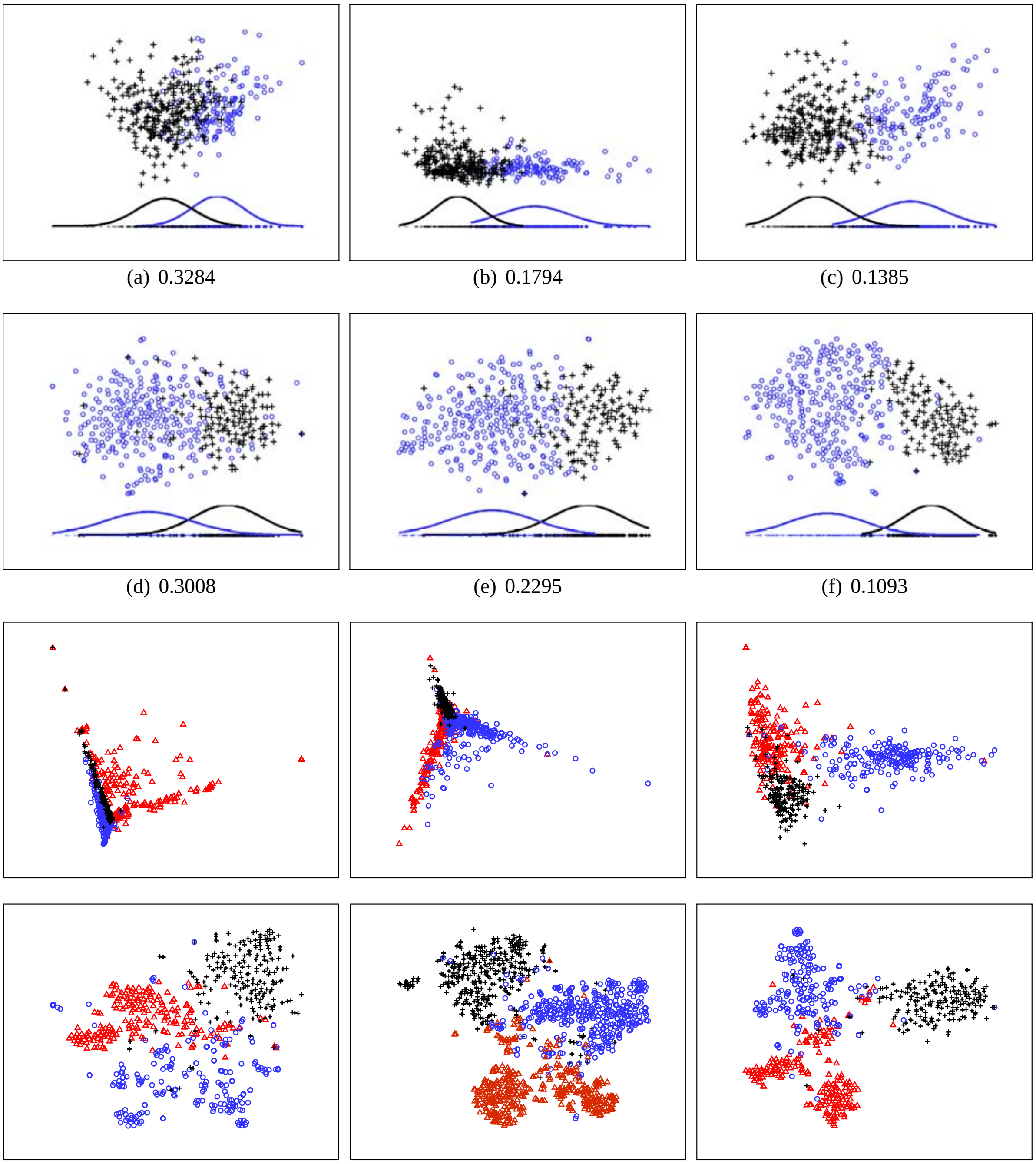
with $\sum \alpha_i = 1$ and $\alpha_i \geq 0$. Table 3 displays quantitative results using evaluation measures (i), (ii) and (iii) where k is chosen to be 5 for (iii). The first four rows correspond to method A, B, C and D and the bottom row corresponds to a convex combination found which minimizes the unsupervised evaluation measure (ii) (i.e. the search for the optimal combination is based on (ii) that does not require labeled data). Note that the convex combination also outperforms method A, B, C, and D for measure (i) and more impressively for measure (iii) which is a supervised measure that uses labeled data. In general, by combining heterogeneous types of domain knowledge, we may further improve the quality of dimensionality reduction for visualization, and the search for such a combination may be accomplished without the use of labeled data.

$(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$	(i)	(ii)	(iii) (k=5)
(1,0,0,0)	0.5756	-3.9334	0.7666
(0,1,0,0)	0.5645	-4.6966	0.7765
(0,0,1,0)	0.5155	-5.0154	0.8146
(0,0,0,1)	0.6035	-3.1154	0.8245
(0.3,0.4,0.1,0.2)	**0.4735**	**-5.1154**	**0.8976**

Table 3: Three evaluation measures (i), (ii), and (iii) (see the beginning of the section for description) for convex combinations (6) using different values of α. The first four rows represent methods A, B, C, and D. The bottom row represents a convex combination whose coefficients were obtained by searching for the minimizer of measure (ii). Interestingly the minimizer also performs well on measure (i) and more impressively on the labeled measure (iii).

Finally, we demonstrate the effect of domain knowledge on a new dataset that consists of all oral papers appearing in ACL 2001 – 2009. For the purpose of manual specification, we obtain 1545 unique words from paper titles, and assign for each word relatedness scores for the following clusters: morphology/phonology, syntax/parsing, semantics, discourse/dialogue, generation/summarization, machine translation, retrieval/categorization and machine learning. The score takes value from 0 to 2, where 2 represents the most relevant. The score information is then used to generate the transformation matrix R. We also assign for each word an importance value ranging from 0 to 3 (larger the value, more important the word). This information is used to generate the diagonal matrix D.

Figure 4 shows the projection of all 2009 papers using t-SNE (papers from 2001 to 2008 are used to estimate contextual diffusion). Using Euclidean geometry $H = I$ (Figure 4 left) results in a Gaussian like distribution which does not provide much insight into the data. Using a manually



(a) 0.3284

(b) 0.1794

(c) 0.1385

(d) 0.3008

(e) 0.2295

(f) 0.1093

Figure 3: Qualitative evaluation of dimensionality reduction for the sentiment domain (top two rows) and the newsgroup domain (bottom two rows). The first and the third rows display PCA reduction while the second and the fourth display t-SNE. The left column correspond to no domain knowledge ($H = I$) reverting PCA and t-SNE to their original form. The middle column corresponds to manual specification (method A). The right column corresponds to contextual diffusion (method B). Different groups (sentiment labels or newsgroup labels) are marked with different colors and marks.
In the sentiment case (top two rows) the graphs were rotated such that the direction returned by applying Fisher linear discriminant onto the projected 2D coordinates aligns with the positive x-axis. The bell curves are Gaussian distributions fitted from the x-coordinates of the projected data points (after rotation). The numbers displayed in each sub-figure are computed from measure (iv).

	Dennis Schwartz		James Berardinelli		Scott Renshaw		Steve Rhodes	
	PCA	t-SNE	PCA	t-SNE	PCA	t-SNE	PCA	t-SNE
$H = I$	1.8625	1.8781	1.4704	1.5909	1.8047	1.9453	1.8013	1.8415
A	1.8474	1.7909	1.3292	1.4406	1.6520	1.8166	**1.4844**	1.6610
B	**1.4254**	**1.5809**	**1.3140**	**1.3276**	**1.5133**	**1.6097**	1.5053	**1.6145**
C	1.6868	1.7766	1.3813	1.4371	1.7200	1.8605	1.7750	1.7979
$H = I$	0.6404	0.7465	0.8481	0.8496	0.6559	0.6821	0.6680	0.7410
A	0.6011	0.7779	**0.9224**	0.8966	0.7424	0.7411	**0.8350**	0.8513
B	**0.8831**	**0.8554**	0.9188	**0.9377**	**0.8215**	**0.8332**	0.8124	**0.8324**
C	0.7238	0.7981	0.8871	0.9093	0.6897	0.7151	0.6724	0.7726

Table 1: Quantitative evaluation of dimensionality reduction for visualization in the sentiment domain. Each of the four columns corresponds to a different movie critic from the Cornell dataset (see text). The top five rows correspond to measure (i) (lower is better) and the bottom five rows correspond to measure (iii) ($k = 5$, higher is better). Results were averaged over 40 cross validation iterations. We conclude that all methods outperform the original $H = I$ with the contextual diffusion and manual specification generally outperforming the others.

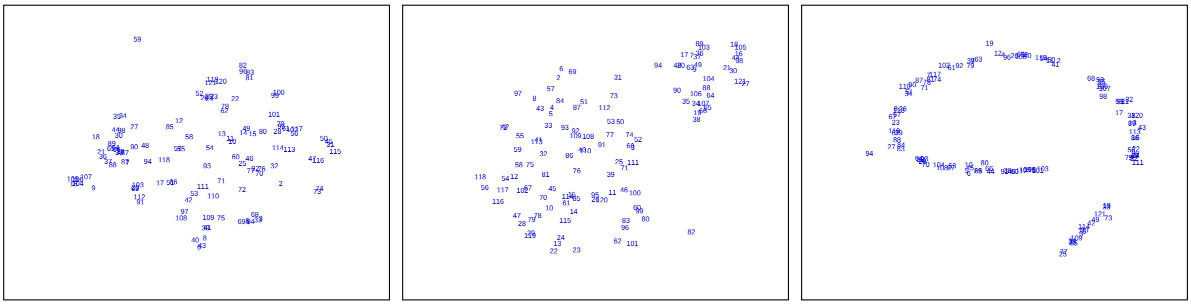


Figure 4: Qualitative evaluation of dimensionality reduction for the ACL dataset using t-SNE. Left: no domain knowledge ($H = I$); Middle: manual specification (method A); Right: contextual diffusion (method B). Each document is labeled by its assigned id from ACL anthology. See text for more details.

specified H (Figure 4 left) we get two clear clusters, the smaller containing papers dealing with machine translation and multilingual tasks. Interestingly, the contextual diffusion results in a one-dimensional manifold. Investigating the papers along the curve (from bottom to top) we find that it starts with papers discussing semantics and discourse (south), continues to structured prediction and segmentation (east), continues to parsing and machine learning (north), and then moves to sentiment prediction, summarization and IR (west) before returning to the center. Another interesting insight that we can derive is the relative discontinuity between the bottom part (semantics and discourse) and the rest of the curve. It seems spatial separability is higher in that area than in the other areas where the curve nicely traverses different regions continuously.

6 Discussion

In this paper we introduce several ways of incorporating domain knowledge into dimensionality reduction for visualizing text documents. The proposed methods all outperform in general the baseline $H = I$, which is the one currently used in most text visualization systems.

The answer to the question of which method is best depends on both the domain and the task at hand. For small tasks with limited vocabulary, manual specification could achieve best results. A large vocabulary size makes manual specification less accurate and effective. In cases where we have access to a large external corpus that is similar to the one we are interested in visualizing, contextual diffusion is an excellent choice. Lacking such a domain specific dataset estimating the contextual distribution using the generic Google n-gram is a good substitute. Word-Net captures relationships (such as synonyms and hyponyms) other than occurrence statistics between vocabulary words, and could be useful for certain tasks. Finally, the effectiveness of dimensionality reduction methods can be increased further by carefully combining different types of domain knowledge ranging from semantic similarity to occurrence statistics.

References

Blei, D., A. Ng, , and M. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Brants, T. and A. Franz. 2006. Web 1T 5-gram Version 1.

Budanitsky, A. and G. Hirst. 2001. Semantic distance in wordnet: An experimental, application-oriented evaluation of five measures. In *NAACL Workshop on WordNet and other Lexical Resources*.

Burges, C. 2009. Dimension reduction: A guided tour. Technical Report MSR-TR-2009-2013, Microsoft Research.

Davies, D. L. and D. W. Bouldin. 2000. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(4):224–227.

Dillon, J., Y. Mao, G. Lebanon, and J. Zhang. 2007. Statistical translation, heat kernels, and expected distances. In *Uncertainty in Artificial Intelligence*, pages 93–100. AUAI Press.

Duda, R. O., P. E. Hart, and D. G. Stork. 2001. *Pattern classification*. Wiley New York.

Havre, S., E. Hetzler, P. Whitney, and L. Nowell. 2002. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1).

Hearst, M. A. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

Jiang, J. J. and D. W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference Research on Computational Linguistics (ROCLING X)*.

Jurafsky, D. and J. H. Martin. 2008. *Speech and Language Processing*. Prentice Hall.

Lafferty, J. and G. Lebanon. 2005. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6:129–163.

Lebanon, G. 2006. Metric learning for text documents. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(4):497–508.

Lewis, D., Y. Yang, T. Rose, and F. Li. 2004. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

Mao, Y., J. Dillon, and G. Lebanon. 2007. Sequential document visualization. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1208–1215.

Paley, W. B. 2002. TextArc: Showing word frequency and distribution in text. In *IEEE Symposium on Information Visualization Poster Compendium*.

Pang, B. and L. Lee. 2004. A sentimental eduction: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proc. of the Association of Computational Linguistics*.

Pang, B., L. Lee, and S. Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proc. of the Conference on Empirical Methods in Natural Language Processing*.

Roweis, S. and L. Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326.

Spoerri, A. 1993. InfoCrystal: A visual tool for information retrieval. In *Proc. of IEEE Visualization*.

Thomas, J. J. and K. A. Cook, editors. 2005. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society.

van der Maaten, L. and G. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Xing, E., A. Ng, M. Jordan, and S. Russel. 2003. Distance metric learning with applications to clustering with side information. In *Advances in Neural Information Processing Systems, 15*.