Learning Nonstructural Distance Metric by Minimum Cluster Distortions

Daichi Mochihashi, Genichiro Kikui
ATR Spoken Language Translation research laboratories
Hikaridai 2-2-2, Keihanna Science City
Kyoto 619-0288, Japan
daichi.mochihashi@atr.jp
genichiro.kikui@atr.jp

Kenji Kita
Center for Advanced Information Technology, Tokushima University
Minamijosanjima 2-1
Tokushima 770-8506, Japan
kita@is.tokushima-u.ac.jp

Abstract

Much natural language processing still depends on the Euclidean (cosine) distance function between two feature vectors, but this has severe problems with regard to feature weightings and feature correlations. To answer these problems, we propose an optimal metric distance that can be used as an alternative to the cosine distance, thus accommodating the two problems at the same time. This metric is optimal in the sense of global quadratic minimization, and can be obtained from the clusters in the training data in a supervised fashion.

We confirmed the effect of the proposed metric distance by a synonymous sentence retrieval task, document retrieval task and the K-means clustering of general vectorial data. The results showed constant improvement over the baseline method of Euclid and tf.idf, and were especially prominent for the sentence retrieval task, showing a 33% increase in the 11-point average precision.

1 Introduction

Natural language processing involves many kinds of linguistic expressions, such as sentences, phrases, documents and the collection of documents. Comparing these expressions based on semantic proximity is a fundamental task and has many applications. Generally, two basic approaches exist to compare two expressions: (a) structural and (b) nonstructural. Structural approaches make use of syntactic parsing or dependency analysis to make a rigorous comparison; nonstructural approaches use vector representation and provide a rough but fast comparison that is required for search/retrieval from a vast amount of corpora. While structural approaches have recently become available in a kernel-based sophisticated treatment (Collins and Duffy, 2001; Suzuki et al., 2003), here we concentrate on nonstructural comparison. This is not only because nonstructural comparison constitutes an integral part in structural methods (that is, even in hierarchical methods the leaf comparison is still atomic), but because it is frequently embedded in many applications where structural parsings are not available or computationally too expensive. For example, information retrieval has long used the 'bag of words' approach (Baeza-Yates and Ribeiro-Neto, 1999; Schütze, 1992) mainly due to a lack of scalable segmentation algorithms and the huge amount of data involved. While segmentation algorithms, such as TEXTTILING (Hearst, 1994) and its recent successors using the inter-paragraph similarity matrix (Choi, 2000), all themselves use nonstructural cosine similarity as a measure of semantic proximity between paragraphs.

However, the distance function so far has been largely defined and used *ad hoc*, usually by a tf.idf weighting scheme (Salton and Yang, 1973) and a simple cosine similarity, equivalently, an Euclidean dot product. In this paper, we propose an optimal distance function that is parameterized by a global metric matrix. This metric is optimal in the sense of global quadratic minimization, and can be learned from the given clusters in the training data. These clusters are often attributable with many forms, such as paragraphs, documents or document collections, as long as the items in the training data are not completely independent.

This paper is organized as follows. In section 2 we describe the issue of traditional Euclidean distances, and section 3 places it into general perspective with related works in machine learning. Section 4 introduces the proposed metric, and section 5 validates its effect on the task of sentence retrieval, document retrieval and the K-means clustering. Sections 6 and 7 present discussions and the conclusion.

2 Issues with Euclidean distances

When we address nonstructural matching, linguistic expressions are often modeled by a feature vector $\vec{x} \in \mathbb{R}^n$, with its elements $x_1 \dots x_n$ corresponding to the number of occurrences of i'th feature. If features are simply words, this is called a 'bag of words'; but in general, features are not restricted to this kind, and we will use the general term "feature"

in the rest of the paper.

To measure the distance between two vectors $\vec{u}, \vec{v}$, a dot product or Euclidean distance

$$d(\vec{u},\vec{v})^2 = (\vec{u}-\vec{v})^T(\vec{u}-\vec{v}) = \sum_{i=1}^{n}(u_i-v_i)^2 \quad (1)$$

(where ^{T} denotes a transposition) has been employed so far [1], with a heuristic feature weighting such as tf.idf in a preprocessing stage.

However, there are two main problems with this distance:

(1) The correlation between features is ignored.
(2) Feature weighting is inevitably arbitrary.

Problem (1) is especially important in languages, because linguistic features (e.g., words) generally have strong correlations between them, such as collocations or typical constructions. But this correlation cannot be considered in a simple dot product. While it is possible to address this with a specific kernel function, such as polynomials (Müller et al., 2001), this is not available for many problems, such as information retrieval or question answering, that do not fit classifications or cannot be easily "kernelized". Problem (2) is a more subtle but inherent one: while tf.idf often works properly in practice, there are several options, especially in tf such as logs or square roots, but we have no principle with which to choose from. Further, it has no theoretical basis that gives any optimality as a distance function.

3 Related Works

The issues above of feature correlations and feature weightings can be summarized as a problem of defining an appropriate metric in the feature space, based on the distribution of data. This problem has recently been highlighted in the field of machine learning research. (Xing et al., 2002) has an objective that is quite similar to that of this paper, and gives a metric matrix that resembles ours based on sample pairs of "similar points" as training data. (Bach and Jordan, 2004) and (Schultz and Joachims, 2004) seek to answer the same problem with an additional scenario of spectral clustering and relative comparisons in Support Vector Machines, respectively. In this aspect, our work is a straight successor of (Xing et al., 2002) where its general usage in vector space is preserved. We offer a discussion on the similarity to our method and our advantages in section 6. Finally, we note that the Fisher kernel of (Jaakkola and Haussler, 1999) has the same concept that gives an appropriate similarity of two data through the Fisher information matrix obtained from the empirical distribution of data. However, it is often approximated by a unit matrix because of its heavy computational demand.

In the field of information retrieval, (Jiang and Berry, 1998) proposes a Riemannian SVD (R-SVD) from the viewpoint of relevance feedback. This work is close in spirit to our work, but is not aimed at defining a permanent distance function and does not utilize cluster structures existent in the training data.

4 Defining an Optimal Metric

To solve the problems in section 2, we note the function that synonymous clusters play. There are many levels of (more or less) synonymous clusters in linguistic data: phrases, sentences, paragraphs, documents, and, in a web environment, the site that contains the document. These kinds of clusters can often be attributed to linguistic expressions because they nest in general so that each expression has a parent cluster.

Since these clusters are synonymous, we can expect the vectors in each cluster to concentrate in the ideal feature space. Based on this property, we can introduce an optimal weighting and correlation in a supervised fashion. We will describe this method below.

4.1 The Basic Idea

As stated above, vectors in the same cluster must have a small distance between each other in the ideal geometry. When we measure an L_2-distance between $\vec{u}$ and $\vec{v}$ by a Mahalanobis distance parameterized by M:

$$d_M(\vec{u},\vec{v})^2 = (\vec{u}-\vec{v})^T M(\vec{u}-\vec{v}) = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}(u_i-v_i)(u_j-v_j), \quad (2)$$

where symmetric metric matrix M gives both corresponding feature weights and feature correlations. When we take $M = I$ (unit matrix), we recover the original Euclidean distance (1).

Equation (2) can be rewritten as (3) because M is symmetric:

$$d_M(\vec{u},\vec{v})^2 = (M^{1/2}(\vec{u}-\vec{v}))^T(M^{1/2}(\vec{u}-\vec{v})). \quad (3)$$

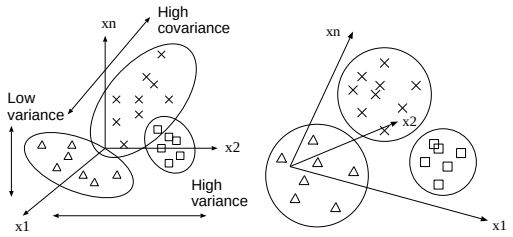
Therefore, this distance amounts to a Euclidean distance in $M^{1/2}$-mapped space (Xing et al., 2002).

Note that this distance is global, and *different* from the ordinary Mahalanobis distance in pattern

[1] When we normalize the length of the vectors $|\vec{u}| = |\vec{v}| = 1$ as commonly adopted, $(\vec{u}-\vec{v})^T(\vec{u}-\vec{v}) = |\vec{u}|^2 + |\vec{v}|^2 - 2\vec{u}\cdot\vec{v} \propto -\vec{u}\cdot\vec{v} = -\cos(\vec{u},\vec{v})$; therefore, this includes a cosine similarity (Manning and Schütze, 1999).

recognition (for example, (Duda et al., 2000)) that is defined for each cluster one by one, using a cluster-specific covariance matrix. That type of distance cannot be generalized to new kinds of data; therefore, it has been used for local classifications. What we want is a global distance metric that is generally useful, not a measure for classification to predefined clusters. In this respect, (Xing et al., 2002) shares the same objective as ours.

Therefore, we require an optimization over all the clusters in the training data. Generally, data in the clusters are distributed as in figure 1(a), comprising ellipsoidal forms that have high (co)variances for some dimensions and low (co)variances for other dimensions. Further, the cluster is not usually aligned to the axes of coordinates. When we find a global metric matrix M that minimizes the cluster distortions, namely, one that reduces high variances and expands low variances for the data to make a spherical form as good as possible in the $M^{1/2}$-mapped space (figure 1(b)), we can expect it to capture necessary and unnecessary variations and correlations on the features, combining information from many clusters to produce a more reliable metric that is not locally optimal. We will find this optimal M below.



(a) Original space (b) Mapped space

Figure 1: Geometry of feature space.

4.2 Global optimization over clusters

Suppose that each data (for example, sentences or documents) is a vector $\vec{s} \in \mathbb{R}^n$, and the whole corpus can be divided into N clusters, $X_1 \dots X_N$. That is, each vector has a dimension n, and the number of clusters is N. For each cluster X_i, cluster centroid c_i is calculated as $\vec{c}_i = 1/|X_i| \sum_{\vec{s} \in X_i} \vec{s}$, where $|X|$ denotes the number of data in X. When necessary, each element in $\vec{s}_j$ or $\vec{c}_i$ is referenced as s_{jk} or c_{ik} $(k = 1 \dots n)$.

The basic idea above is formulated as follows. We seek the metric matrix M that minimizes the distance between each data $\vec{s}_j$ and the cluster centroid $\vec{c}_i$, $d_M(\vec{s}_j, \vec{c}_i)$ for all clusters $X_1 \dots X_N$. Mathematically, this is formulated as a quadratic minimization problem

$$M = \arg\min_M \sum_{i=1}^{N} \sum_{\vec{s}_j \in X_i} d_M(\vec{s}_j, \vec{c}_i)^2$$
$$= \arg\min_M \sum_{i=1}^{N} \sum_{\vec{s}_j \in X_i} (\vec{s}_j - \vec{c}_i)^T M (\vec{s}_j - \vec{c}_i) \quad (4)$$

under a scale constraint ($|\cdot|$ means determinant)

$$|M| = 1. \quad (5)$$

Scale constraint (5) is necessary for excluding a degenerate solution $M = O$. 1 is an arbitrary constant: when we replace 1 by c, $c^2 M$ becomes a new solution. This minimization problem is an extension to the method of *MindReader* (Ishikawa et al., 1998) to multiple clusters, and has a unique solution below.

Theorem The matrix that solves the minimization problem (4,5) is

$$M = |A|^{1/n} A^{-1}, \quad (6)$$

where $A = [a_{kl}]$ is defined by

$$a_{kl} = \sum_{i=1}^{N} \sum_{s_j \in X_i} (s_{jl} - c_{il})(s_{jk} - c_{ik}) \,. \quad (7)$$

Proof: See Appendix A. ■

When A is singular, we can use as A^{-1} a Moore-Penrose matrix pseudoinverse A^+. Generally, A consists of linguistic features and is very sparse, and often singular. Therefore, A^+ is nearly always necessary for the above computation. For details, see Appendix B.

4.3 Generalization

While we assumed through the above construction that each cluster is equally important, this is not the case in general. For example, clusters with a small number of data may be considered weak, and in the hierarchical clustering situation, a "grandmother" cluster may be weaker. If we have confidences $\xi_1 \dots \xi_N$ for the strength of clustering for each cluster $X_1 \dots X_N$, this information can be incorporated into (4) by a set of normalized cluster weights ξ_i^*:

$$M = \arg\min_M \sum_{i=1}^{N} \xi_i^* \sum_{\vec{s}_j \in X_i} (\vec{s}_j - \vec{c}_i)^T M (\vec{s}_j - \vec{c}_i),$$

where $\xi_i^* = \xi_i / \sum_{j=1}^{N} \xi_j$, and we obtain a respectively weighted solution in (7). Further, we note that when $N = 1$, this metric recovers the ordinary Mahalanobis distance in pattern recognition. However, we used equal weights for the experiments below because the number of data in each cluster was approximately equal.

5 Experiments

We evaluated our metric distance on the three tasks of synonymous sentence retrieval, document retrieval, and the K-means clustering of general vectorial data. After calculating M on the training data of clusters, we applied it to the test data to see how well its clusters could be recovered. As a measure of cluster recovery, we use 11-point average precision and R-precision for the distribution of items of the same cluster in each retrieval result. Here, R equals the cardinality of the cluster; therefore, R-precision shows the precision of cluster recovery.

5.1 Synonymous sentence retrieval

5.1.1 Sentence cluster corpus

We used a paraphrasing corpus of travel conversations (Sugaya et al., 2002) for sentence retrieval. This corpus consists of 33,723,164 Japanese translations, each of which corresponds to one of the original English sentences. By way of this correspondence, Japanese sentences are divided into 10,610 clusters. Therefore, each cluster consists of Japanese sentences that are possible translations from the same English seed sentence that the cluster has. From this corpus, we constructed 10 sets of data. Each set contains random selection of 200 training clusters and 50 test clusters, and each cluster contains a maximum of 100 sentences [2]. Experiments were conducted on these 10 datasets for each level of dimensionality reduction (see below) to produce average statistics.

5.1.2 Features and dimensionality reduction

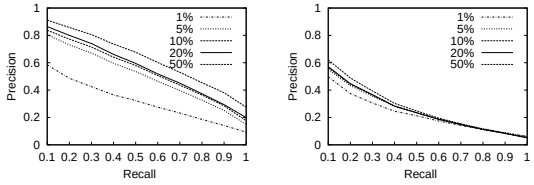
As a feature of a sentence, we adopted unigrams of all words and bigrams of functional words from the part-of-speech tags, because the sequence of functional words is important in the conversational corpus.

While the lexicon is limited for travel conversations, the number of features exceeds several thousand or more. This may be prohibitive for the calculation of the metric matrix, therefore, we additionally compressed the features with SVD, the same method used in Latent Semantic Indexing (Deerwester et al., 1990).

5.1.3 Sentence retrieval results

Qualitative result Figure 5 (last page) shows a sample retrieval result. A sentence with (*) mark at the end is the correct answer, that is, a sentence from the same original cluster as the query. We can see that the results with the metric distance contain less noise than a standard Euclid baseline with tf.idf weighting, achieving a high-precision retrieval. Although the high rate of dimensionality reduction in figure 6 shows degradation due to the dimension contamination, the effect of metric distance is still apparent despite bad conditions.

Quantitative result Figure 2 shows the averaged precision-recall curves of retrieval and figure 3 shows 11-point average precisions, for each rate of dimensionality reduction. Clearly, our method achieves higher precision than the standard method, and does not degrade much with feature compressions unless we reduce the dimension too much, i.e., to $< 5\%$.



(a) Metric distance + idf (b) Euclidean + idf

Figure 2: Precision-recall of sentence retrieval.

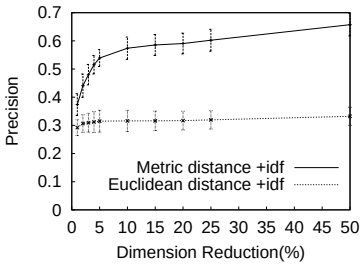


Figure 3: 11-point average precision.

5.2 Document retrieval

As a method of tackling clusters of texts, the text classification task has recently made great advances with a Naïve Bayes or SVM classifiers (for example, (Joachims, 1998)). However, they all aim at classifying texts into a few predefined clusters, and cannot deal with a document that fits neither of the clusters. For example, when we regard a website as a cluster of documents, the possible clusters are numerous and constantly increasing, which precludes classificatory approaches. For these circumstances, document clustering or retrieval will benefit from a global distance metric that exploits the multitude of cluster structures themselves.

5.2.1 Newsgroup text dataset

For this purpose, we used the 20-Newsgroup dataset (Lang, 1995). This is a standard text classification dataset that has a relatively large number of classes,

[2] When the number of data in the cluster exceeds this limit, 100 sentences are randomly sampled. All sampling are made without replacement.

20. Among the 20 newsgroups, we selected 16 clusters of training data and 4 clusters of test data, and performed 5-fold cross validation. The maximum number of documents per cluster is 100, and when it exceeds this limit, we made a random sampling of 100 documents as the sentence retrieval experiment.

Because our proposed metric is calculated from the distribution of vectors in high-dimensional feature space, it becomes inappropriate if the norm of the vectors (largely proportional to document length) differs much from document to document. [3] Therefore, we used subsampling/oversampling to form a median length (130 words) on training documents. Further, we preprocessed them with tf.idf as a baseline method.

5.2.2 Results

Table 1 shows R-precision and 11-point average precision. Since the test data contains 4 clusters, the baselines of precision are 0.25. We can see from both results that metric distance produces a better retrieval over the tf.idf and dot product. However, refinements in precision are certain (average p = 0.0243) but subtle.

This can be thought of as the effect of the dimensionality reduction performed. We first decompose data matrix X by SVD: $X = USV^{-1}$ and build a k-dimensional compressed representation $X_k = V_k X$; where V_k denotes a k-largest submatrix of V. From the equation (3), this means a Euclidean distance of $M^{1/2}X_k = M^{1/2}V_k X$. Therefore, V_k may subsume the effect of M in a preprocessing stage. Close inspection of table 1 shows this effect as a tradeoff between M and V_k. To make the most of metric distance, we should consider metric induction and dimensionality reduction simultaneously, or reconsider the problem in kernel Hilbert space.

Dim.	R-precision		11-pt Avr. Prec.	
Red.	Metric	Euclid	Metric	Euclid
0.5%	**0.421**	0.399	**0.476**	0.455
1%	**0.388**	0.368	**0.450**	0.430
2%	**0.359**	0.343	**0.425**	0.409
3%	**0.344**	0.330	**0.411**	0.399
4%	**0.335**	0.323	**0.402**	0.392
5%	**0.329**	0.318	**0.397**	0.388
10%	**0.316**	0.307	**0.379**	0.376
20%	**0.343**	0.297	**0.397**	0.365

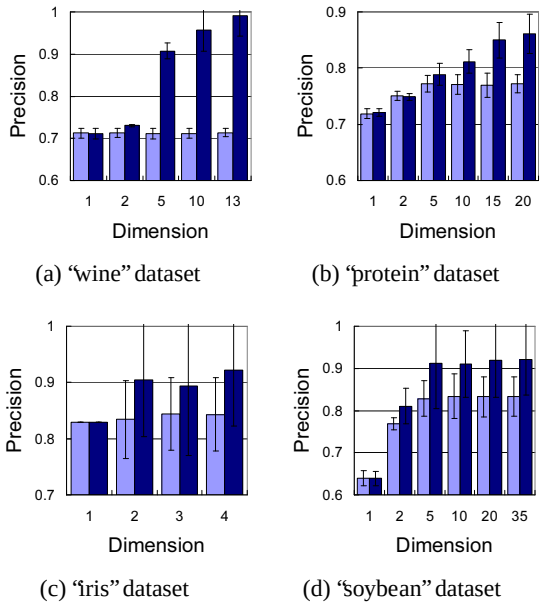
Table 1: Newsgroup text retrieval results.

[3] Normalizing documents to unit length effectively maps them to a high-dimensional hypersphere; this proved to produce an unsatisfactory result. Defining metrics that work on a hypersphere like spherical K-means (Dhillon and Modha, 2001) requires further research.

5.3 K-means clustering and general vectorial data

Metric distance can also be used for clustering or general vectorial data. Figure 4 shows the K-means clustering result of applying our metric distance to some of the UCI Machine Learning datasets (Blake and Merz, 1998). K-means clustering was conducted 100 times with a random start, where K equals the known number of classes in the data [4]. Clustering precision was measured as an average probability that a randomly picked pair of data will conform to the true clustering (Xing et al., 2002).

We also conducted the same clustering for documents of the 20-Newsgroup dataset to get a small increase in precision like the document retrieval experiment in section 5.2.



(a) "wine" dataset (b) "protein" dataset

(c) "iris" dataset (d) "soybean" dataset

Figure 4: K-means clustering of UCI Machine Learning dataset results. The horizontal axis shows compressed dimensions (rightmost is original). The right bar shows clustering precision using Metric distance, and the left bar shows that using Euclidean distance.

6 Discussion

In this paper, we proposed an optimal distance metric based on the idea of minimum cluster distortion in training data. Although vector distances have frequently been used in natural language processing, this is a rather neglected but recently highlighted problem. Unlike recently proposed methods with spectral methods or SVMs, our method assumes no such additional scenarios and can be considered as

[4] Because of the small size of the dataset, we did not apply cross-validation as in other experiments.

a straight successor to (Xing et al., 2002)'s work. Their work has the same perspective as ours, and they calculate a metric matrix A that is similar to ours based on a set S of vector pairs $(\vec{x}_i, \vec{x}_j)$ that can be regarded as similar. They report that the effectiveness of A increases as the number of the training pairs S increases; this requires $O(n^2)$ sample points from n training data, and must be optimized by a computationally expensive Newton-Raphson iteration. On the other hand, our method uses only linear algebra, and can induce an ideal metric using all the training data at the same time. We believe this metric can be useful for many vector-based language processing methods that have used cosine similarity.

There remains some future directions for research. First, as we stated in section 4.3, the effect of a cluster weighted generalized metric must be investigated and optimal weighting must be induced. Second, as noted in section 5.2.1, the dimensionality reduction required for linguistic data may constrain the performance of the metric distance. To alleviate this problem, simultaneous dimensionality reduction and metric induction may be necessary, or the same idea in a kernel-based approach is worth considering. The latter obviates the problem of dimensionality, while it restricts the usage to a situation where the kernel-based approach is available.

7 Conclusion

We proposed a global metric distance that is useful for clustering or retrieval where Euclidean distance has been used. This distance is optimal in the sense of quadratic minimization over all the clusters in the training data. Experiments on sentence retrieval, document retrieval and K-means clustering all showed improvements over Euclidean distance, with a significant refinement with tight training clusters in sentence retrieval.

Acknowledgement

The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology entitled "A study of speech dialogue translation technology based on a large corpus".

References

Francis R. Bach and Michael I. Jordan. 2004. Learning Spectral Clustering. In *Advances in Neural Information Processing Systems 16*. MIT Press.

Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press / Addison-Wesley.

C. L. Blake and C. J. Merz. 1998. UCI Repository of machine learning databases. http://www.ics.uci.edu/˜mlearn/MLRepository.html.

Freddy Y. Y. Choi. 2000. Advances in domain independent linear text segmentation. In *Proceedings of NAACL-00*.

Michael Collins and Nigel Duffy. 2001. Convolution Kernels for Natural Language. In *NIPS 2001*.

S. Deerwester, Susan T. Dumais, and George W. Furnas. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Inderjit S. Dhillon and Dharmendra S. Modha. 2001. Concept Decompositions for Large Sparse Text Data Using Clustering. *Machine Learning*, 42(1/2):143–175.

Richard O. Duda, Peter E. Hart, and David G. Stork. 2000. *Pattern Classification *Second Edition*. John Wiley & Sons.

Marti Hearst. 1994. Multi-paragraph segmentation of expository text. In *32nd. Annual Meeting of the Association for Computational Linguistics*, pages 9–16.

Yoshiharu Ishikawa, Ravishankar Subramanya, and Christos Faloutsos. 1998. MindReader: Querying Databases Through Multiple Examples. In *Proc. 24th Int. Conf. Very Large Data Bases*, pages 218–227.

Tommi S. Jaakkola and David Haussler. 1999. Exploiting generative models in discriminative classifiers. In *Proc. of the 1998 Conference on Advances in Neural Information Processing Systems*, pages 487–493.

Eric P. Jiang and Michael W. Berry. 1998. Information Filtering Using the Riemannian SVD (R-SVD). In *Proc. of IRREGULAR '98*, pages 386–395.

Thorsten Joachims. 1998. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98*, number 1398, pages 137–142.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press.

K. R. Müller, S. Mika, G. Ratsch, and K. Tsuda. 2001. An introduction to kernel-based learning

algorithms. *IEEE Neural Networks*, 12(2):181–201.

G. Salton and C. S. Yang. 1973. On the specification of term values in automatic indexing. *Journal of Documentation*, 29:351–372.

Matthew Schultz and Thorsten Joachims. 2004. Learning a Distance Metric from Relative Comparisons. In *Advances in Neural Information Processing Systems 16*. MIT Press.

Hinrich Schütze. 1992. Dimensions of Meaning. In *Proceedings of Supercomputing'92*, pages 787–796.

F. Sugaya, T. Takezawa, G. Kikui, and S. Yamamoto. 2002. Proposal for a very-large-corpus acquisition method by cell-formed registration. In *Proc. LREC-2002*, volume I, pages 326–328.

Jun Suzuki, Tsutomu Hirao, Yutaka Sasaki, and Eisaku Maeda. 2003. Hierarchical Directed Acyclic Graph Kernel: Methods for Structured Natural Language Data. In *Proc. of the 41th Annual Meeting of Association for Computational Linguistics (ACL2003)*, pages 32–39.

Eric W. Weisstein. 2004. Moore-Penrose Matrix Inverse. http://mathworld.wolfram.com/Moore-PenroseMatrixInverse.html.

Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, and Stuart Russell. 2002. Distance metric learning, with application to clustering with side-information. In *NIPS 2002*.

Appendix A. Derivation of the metric matrix

Here we prove theorem 1, namely deriving M that satisfies the condition

$$\min_{M} \sum_{i=1}^{n} \sum_{\vec{s}_j \in X_i} (\vec{s}_j - \vec{c}_i)^T M (\vec{s}_j - \vec{c}_i) \,, \qquad (8)$$

under the constraint

$$|M| = 1. \qquad (9)$$

Expanding (8), we get

$$\sum_i \sum_{\vec{s}_j} \left[\sum_{k=1}^{n} \sum_{l=1}^{n} (s_{jk} - c_{ik}) m_{kl} (s_{jl} - c_{il}) \right] \,, \quad (10)$$

and from (9), for all k

$$\sum_{l=1}^{n} (-1)^{k+l} m_{kl} |M_{kl}| = 1 \,.$$

Therefore

$$\sum_{k=1}^{n} \sum_{l=1}^{n} (-1)^{k+l} m_{kl} |M_{kl}| = n, \qquad (11)$$

where M_{kl} denotes an adjugate matrix of m_{kl}.

Therefore, we come to minimize (10) under the constraint (11).

By introducing the Lagrange multiplier λ, we define

$$L = \sum_{i=1}^{N} \sum_{\vec{s}_j} \left[\sum_k \sum_l (s_{jk} - c_{ik}) m_{kl} (s_{jl} - c_{il}) \right] - \lambda \left[\sum_k \sum_l (-1)^{k+l} m_{kl} |M_{kl}| - n \right].$$

Differentiating by m_{kl} and setting to zero, we obtain

$$\begin{aligned} \frac{\partial L}{\partial m_{kl}} &= \sum_i \sum_{\vec{s}_j} (s_{jk} - c_{ik})(s_{jl} - c_{il}) \\ &\quad - \lambda (-1)^{k+l} |M_{kl}| = 0 \\ \Leftrightarrow \; |M_{kl}| &= \frac{\sum_i \sum_{\vec{s}_j} (s_{jk} - c_{ik})(s_{jl} - c_{il})}{\lambda (-1)^{k+l}}. \end{aligned} \qquad (12)$$

Let us define $M^{-1} = [m_{kl}^{-1}]$. Then,

$$\begin{aligned} m_{kl}^{-1} &= \frac{(-1)^{k+l} |M_{kl}|}{|M|} \\ &= (-1)^{k+l} |M_{kl}| \quad (\because (9)) \\ &= \frac{\sum_i \sum_{\vec{s}_j} (s_{jk} - c_{ik})(s_{jl} - c_{il})}{\lambda} \\ &\quad (\because (12)) \end{aligned} \qquad (13)$$

Therefore, when we define

$$A = [a_{kl}] \qquad (14)$$

as

$$a_{kl} = \sum_{i=1}^{N} \sum_{\vec{s}_j \in X_i} (s_{jl} - c_{il})(s_{jk} - c_{ik}) \,, \qquad (15)$$

from (13),

$$\begin{aligned} A &= \lambda M^{-1} \\ \therefore |A| &= \lambda^n |M^{-1}| = \lambda^n \\ \therefore \; \lambda &= |A|^{1/n} \,, \end{aligned}$$

where A is defined by (14), (15). ■

Appendix B.
Moore-Penrose Matrix Pseudoinverse

The Moore-Penrose matrix pseudoinverse A^+ of A is a unique matrix that has a property of normal inverse in that $x = A^+y$ is a shortest length least squares solution to $Ax = y$ even if A is singular (Weisstein, 2004).

A^+ can be calculated simply by a MATLAB function `pinv`. Or alternatively (Ishikawa et al., 1998), we can decompose A as

$$A = U\Sigma U^T ,$$

where U is an orthonormal $n \times n$ matrix and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_R, 0, \ldots, 0)$ $(R = \mathrm{rank}(A))$. Then, A^+ is calculated as

$$A^+ = U\Sigma^+ U^T ,$$

where $\Sigma^+ = \mathrm{diag}(1/\sigma_1, \ldots, 1/\sigma_R, 0, \ldots, 0)$. Therefore,

$$M = (\sigma_1\sigma_2 \cdots \sigma_R)^{1/R} A^+. \quad \blacksquare$$

Query: “合計でいくらですか”
(‘How much is the total?’)

Metric distance:

distance	synonymous sentence
0.2712	合計でいくらでしょうか *
0.3444	内金はいくらですか
0.3444	入場料はいくらですか
0.369	手付金はいくらですか
0.4377	合計でいくらいたしますか *
0.4479	合計でいくらいたしますでしょうか *
0.4505	全部でいくらですか *
0.4558	合計でいくらになりますか *
0.4602	合計でいくらになりますでしょうか *
0.4682	合計でいくらになるでしょうか *
0.4729	合計でいくらしますか *
0.4851	合計でいくらしますでしょうか *

Euclidean distance:

distance	synonymous sentence
0.1732	全部でいくらですか *
1.781	合計でおいくらですか *
1.902	紫外線防止ですか
1.966	内金はいくらですか
1.966	入場料はいくらですか
1.974	手付金はいくらですか
1.983	全部でおいくらですか *
2.283	どんな兆候ですか
2.505	どんな症状ですか
2.65	お一人ですか
2.729	放送で呼び出してください
2.749	紫外線防止ですね

(* denotes the right answers.)

Figure 5: Sentence retrieval example.

Query: “デザートに果物をくれないでしょうか”
(‘I’d like some fruit for dessert.’)

Metric distance:

distance	synonymous sentence
0.3531	請求書をすぐにくれないでしょうか
0.3709	デザートとして果物をくれますか *
0.596	請求書をすぐにくれませんか
0.6104	伝票をすぐにくれますか
0.621	伝票をすぐにくれますでしょうか
0.6255	お勘定書をすぐにくれますか
0.6295	伝票をすぐにくれませんでしょうか
0.6343	お勘定書をすぐにくれませんか
0.6685	伝票をすぐにくれないですか
0.7966	デザートには果物をくれないですか *

Euclidean distance:

distance	synonymous sentence
1.036	請求書をすぐにくれないでしょうか
1.421	朝ごはんを部屋に運んでもらえないでしょうか
1.491	ウィスキーを二人分くれないでしょうか
1.499	ウィスキーを二つくれないでしょうか
1.535	薬をくれないでしょうか
1.622	朝食を部屋に運んでもらえないでしょうか
1.622	朝食を部屋に運んでもらえないでしょうか
:	:
2.787	デザートとして何か果物をくれないでしょうか *
2.854	この円をポンドに換算くださらないでしょうか

Figure 6: High rate of dimensionality reduction.