Improving Document Clustering Using Automated Machine Translation

Xiang Wang
Department of Computer Science
University of California, Davis
xiang@ucdavis.edu

Buyue Qian
Department of Computer Science
University of California, Davis
byqian@ucdavis.edu

Ian Davidson
Department of Computer Science
University of California, Davis
davidson@cs.ucdavis.edu

ABSTRACT

With the development of statistical machine translation, we have ready-to-use tools that can translate documents in one language into many different languages. These translations provides different yet correlated views of the same set of documents. This gives rise to a natural question: can we use the extra information to achieve a better clustering of the documents? Some recent work on multiview clustering provided positive answers to this question. In this work, we propose an alternative approach to address this problem using the constrained clustering framework. Unlike traditional Must-Link and Cannot-Link constraints, the constraints generated by machine translation are dense yet noisy. We show how to incorporate this type of constraints by presenting two algorithms, one parametric and one non-parametric. Our algorithms are easy to implement, efficient, and can consistently improve the clustering of real-world data, namely the Reuters RCV1/RCV2 Multilingual Dataset. In contrast to the existing multiview clustering techniques, our technique does not rely on the compatibility and conditional independence assumptions, nor does it involve subtle parameter tuning.

1. INTRODUCTION

1.1 Motivation

Automated Machine Translation (MT) [14] allows documents written in one language to be translated into other languages at very low cost. The field has made great strides recently and many online tools, such as Google Translate, have been made available. This gives rise to a natural question: *can machine translation help us to achieve better clustering of documents?*

This problem has been recently explored from the multiview learning perspective [2, 12, 13] and positive results have been reported. In the multiview learning framework, each original document and its translation are treated as two views of the same data object. Multiview learning makes the assumption that the two views are compatible and conditionally independent [5], thus combining them will lead to a better classification or clustering. However, in reality there is no principled way to check the validity of these assumptions. Another limitation of the existing multiview clustering techniques is that their performance is sensitive to parameter tuning, which relies heavily on the prior knowledge on the relative quality of each particular view.

An alternative approach to incorporating machine translation into document clustering we shall explore is constrained clustering [3]. The basic idea of constrained clustering is to convert side information into Must-Link and Cannot-Link pairwise constraints; then the constraints are enforced on the original dataset to help improving the clustering. Traditionally, the constraints are assumed to be **accurate but sparse**: "accurate" means that they come from domain experts or ground truth, thus they are definite and correct; "sparse" means only a small amount of constraints are incorporated given acquiring them in practice is costly.

However, the constraints generated from machine translation are **dense but noisy**: "dense" means that we have access to large amounts of constraints with no cost or very low cost; "noisy" means that the constraints are noisy and may not necessarily reflect the ground truth clustering. As a result, existing constrained clustering algorithms cannot be directly applied to our problem. A new algorithm is needed to properly incorporate this type of constraints so that it can 1) fully exploit the useful information in the massive constraint set; 2) ignore constraints that are inaccurate and excessive; and 3) not be easily over-constrained as most constrained clustering algorithms are [7].

1.2 Our Contribution

We show how to incorporate massive amounts of noisy side information into spectral clustering and demonstrate that this approach outperforms multiview techniques for MT aided document clustering. We choose a spectral formulation for two reasons: 1) spectral clustering has been proven effective on high-dimensional data, such as text [4], and 2) by converting the two views into two graphs with the same set of nodes we can knowledge transfer between two heterogeneous feature spaces (languages) in a principled manner. Our objective function extends earlier work by Wang and Davidson [18]. Their work was limited to clusterings for $K = 2$ and used a sparse constraint matrix generated from domain experts or ground truth labeling. In our problem setting, the constraint matrix is generated from another graph. As a result, the constraint matrix we use is guar-

anteed to be positive semi-definite (PSD). This difference enables us to extend the formulation in [18] from 2-way partition to K-way partition, and to propose a parametric as well as a non-parametric solution. As compared to existing work in MT aided document clustering, our algorithm has several unique benefits:

- As compared to the multiview techniques, the effectiveness of our algorithm does not rely on the compatibility or the conditional independence assumptions.
- It is easy to implement, efficient, and produces deterministic output.
- It does not require prior knowledge on the dataset, nor does it involve complicated parameter tuning.

Our contributions to the field is to propose a spectral clustering formulation that can handle dense but noisy constraints and we show:

1. Our method is able to improve the clustering on the Reuters RCV1/RCV2 Multilingual Dataset. The improvement is consistent and significant (99% confidence level, see Figure 2 and 3).
2. Our approach outperforms many existing multiview spectral clustering techniques (see Figure 2 and 3).
3. Our approach yields improvements in the vast majority of trials with randomly sampled documents (see Figure 4). This is pragmatically very important given most practitioners only have one dataset (sample) to work with and **average** performance gain is not sufficient.
4. Our algorithm can be extended to other applications and datasets where the side information is dense and not always accurate.

The remainder of the paper is organized as follows: In Section 2 we provide some background knowledge on spectral clustering and the constrained spectral clustering formulation proposed by [18]; we present our algorithm in Section 3; we test our algorithm on real data in Section 4, with comparison to existing techniques; related work is discussed in Section 5; we conclude the paper in Section 6 and discuss future directions.

2. PRELIMINARIES

To make this paper self-contained, we provide some background knowledge and also introduce notations that will be used throughout the rest of the paper (also summarized in Table 1).

2.1 Spectral Clustering

We first introduce the formulation and notation for spectral clustering. Readers who are familiar with the topic can skip to Section 2.2.

Given a graph $\mathcal{G}$ with N nodes, A is the affinity matrix of $\mathcal{G}$. A is symmetric and nonnegative. D is the degree matrix of $\mathcal{G}$:

$$D_{ij} = \begin{cases} \sum_{k=1}^{N} A_{ik} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

Table 1: Table of notations

Notation	Meaning
$\mathcal{G}$	An undirected (weighted) graph
A	The affinity matrix
D	The degree matrix
I	The identity matrix
$\bar{L}$	The normalized graph Laplacian
$\bar{Q}$	The normalized constraint matrix
K	The number of clusters
N	The number of nodes
$\mathbf{v}, V$	The relaxed cluster indicator vector(s)

$L = D - A$ is the graph Laplacian of $\mathcal{G}$, and $\bar{L} = D^{-1/2} L D^{-1/2}$ is called the *normalized* graph Laplacian [17].

The objective function for the normalized min-cut problem is:

$$\begin{aligned} \underset{\mathbf{v} \in \mathbb{R}^N}{\text{argmin}} \quad & \mathbf{v}^T \bar{L} \mathbf{v}, \\ \text{s.t.} \quad & \mathbf{v}^T \mathbf{v} = 1, \ \mathbf{v} \perp D^{1/2} \mathbf{1}. \end{aligned} \tag{1}$$

Shi and Malik [15] showed that the optimal solution to Eq.(1) is the second smallest eigenvector of $\bar{L}$.

For K-way partition, the objective becomes

$$\begin{aligned} \underset{V \in \mathbb{R}^{N \times K}}{\text{argmin}} \quad & tr(V^T \bar{L} V), \\ \text{s.t.} \quad & V^T V = I. \end{aligned} \tag{2}$$

tr is the matrix trace. The optimal solution to Eq.(2) is the top-K smallest eigenvectors of $\bar{L}$, and the clustering assignment is derived from applying K-means to the rows of V [17].

2.2 Constrained Spectral Clustering

Next we briefly summarize the constrained spectral clustering formulation proposed in [18].

Let $Q \in \mathbb{R}^{N \times N}$ be a relaxed constraint matrix. Q is symmetric and

$$Q_{ij} \begin{cases} > 0 & i \text{ and } j \text{ belong to the same cluster} \\ < 0 & i \text{ and } j \text{ belong to different clusters} \\ 0 & \text{unknown} \end{cases}.$$

Given a cut $\mathbf{v}$, the objective is to minimize its cost on $\bar{L}$ and to lower bound its satisfaction on Q with a parameter α. Specifically:

$$\begin{aligned} \underset{\mathbf{v} \in \mathbb{R}^N}{\text{argmin}} \quad & \mathbf{v}^T \bar{L} \mathbf{v}, \\ \text{s.t.} \quad & \mathbf{v}^T \bar{Q} \mathbf{v} \geq \alpha, \\ & \mathbf{v}^T \mathbf{v} = 1, \ \mathbf{v} \perp_{\bar{L}} D^{1/2} \mathbf{1}. \end{aligned} \tag{3}$$

Eq.(3) can be solved by introduce Karush-Kuhn-Tucker conditions [**?**]. The solution is among the eigenvectors of the following the generalized eigenvalue problem:

$$\bar{L} \mathbf{v} = \lambda (\bar{Q} - \alpha I) \mathbf{v}.$$

After removing all negative eigenvectors (which fail to satisfy the lower bound α), the one that minimizes $\mathbf{v}^T \bar{L} \mathbf{v}$ is the solution.

This formulation has several limitations: 1) it is for 2-way clustering instead of K-way; 2) setting the cut-off threshold

Algorithm 1: The parametric version of our algorithm (`csp-p`)

Input: $\bar{L}$, $\bar{Q}$, α, K;
Output: $\mathbf{u}$;
Solve the generalized eigenvalue problem $\bar{L}\mathbf{v} = \lambda\bar{Q}\mathbf{v}$;
Let $\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^N$ be the set of all generalized eigenvectors;
for $i = 1$ **to** N **do**
  if $\mathbf{v}_i^T\bar{Q}\mathbf{v}_i < \alpha$ **then**
    Remove $\mathbf{v}_i$ from $\mathcal{V}$;
  end
end
$V \leftarrow [\,]$;
for $i = 1$ **to** K **do**
  $\mathbf{v}^* = \text{argmin}_{\mathbf{v}\in\mathcal{V}}\, \mathbf{v}^T\bar{L}\mathbf{v}$;
  Remove $\mathbf{v}^*$ from $\mathcal{V}$;
  $V \leftarrow [V, \mathbf{v}^*]$;
end
return $\mathbf{u} \leftarrow Kmeans(V, K)$;

Algorithm 2: The non-parametric version of our algorithm (`csp-n`)

Input: $\bar{L}$, $\bar{Q}$, K;
Output: $\mathbf{u}$;
Solve the generalized eigenvalue problem $\bar{L}\mathbf{v} = \lambda\bar{Q}\mathbf{v}$;
Let $\mathcal{V} = \{\mathbf{v}_i\}_{i=1}^N$ be the set of all generalized eigenvectors;
$V \leftarrow [\,]$;
for $i = 1$ **to** K **do**
  $\mathbf{v}^* = \text{argmin}_{\mathbf{v}\in\mathcal{V}} \frac{\mathbf{v}^T\bar{L}\mathbf{v}}{\mathbf{v}^T\bar{Q}\mathbf{v}}$;
  Remove $\mathbf{v}^*$ from $\mathcal{V}$;
  $V \leftarrow [V, \mathbf{v}^*]$;
end
return $\mathbf{u} \leftarrow Kmeans(V, K)$;

α requires prior knowledge; and 3) it is unclear if it can handle huge amounts of constraints. In Section 3, we will adapt this formulation to our problem and address these limitations.

3. OUR ALGORITHM

In this section, we present our constrained spectral clustering algorithm for document clustering using automated machine translation. The objective function we use is derived from Eq.(3). Unlike earlier work, since Q is obtained from a distance metric, it is guaranteed to be positive semi-definite (PSD). We first show how to construct the graph Laplacian and the constraint matrix in the document clustering setting; then we extend Eq.(3) from 2-way partition to K-way partition (Eq.(4)); we further develop a non-parametric version of the formulation (Eq.6); efficient solutions are provided for both objectives (Algorithm 1 and 2).

3.1 Graph and Constraint Construction

Given a set of N documents, after standard tf-idf indexing and normalization, they can be represented by a set of d-dimensional vectors: $\{\mathbf{x}_i \in \mathbb{R}^d : \|\mathbf{x}_i\| = 1, i = 1, \ldots, N\}$. Let $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$, then

$$A = X^T X$$

is an $N \times N$ cosine similarity matrix and A is positive semi-definite.

In our problem setting, given a set of documents and their translation, we can construct two similarity matrices: $A^{(1)}$ and $A^{(2)}$. We use $A^{(1)}$ as the affinity matrix of the graph, which is equivalent to A in the constrained spectral clustering formulation. We use $A^{(2)}$ as the constraint matrix, which is equivalent to Q in the constrained spectral clustering formulation. The interpretation of $A^{(2)}$ is: the greater $A^{(2)}_{ij}$ is, the more likely document i and j are considered to belong to the same cluster.

Let $D^{(i)}$ be the degree matrix of $A^{(i)}$, $i = 1, 2$. We have:

$$\bar{L} \triangleq I - (D^{(1)})^{-1/2}A^{(1)}(D^{(1)})^{-1/2},$$
$$\bar{Q} \triangleq (D^{(2)})^{-1/2}A^{(2)}(D^{(2)})^{-1/2}.$$

Note that as long as $A^{(1)}$ and $A^{(2)}$ are guaranteed to be PSD, other similarity functions and preprocessing techniques can be freely used. For example, instead of cosine similarity, we can use a Gaussian kernel. We can also apply various dimensionality reduction techniques to improve the quality of the raw data.

3.2 Objective Function

Given $\bar{L}$ and $\bar{Q}$, we first present a parametric formulation for constrained spectral clustering. It is a natural extension of Eq.(3) from 2-way to K-way partition.

$$\begin{aligned} \underset{\mathbf{v}_i\in\mathbb{R}^N|_{i=1}^K}{\text{argmin}} \quad & \sum_{i=1}^K \mathbf{v}_i^T\bar{L}\mathbf{v}_i, \\ \text{s.t.} \quad & \mathbf{v}_i^T\bar{Q}\mathbf{v}_i \geq \alpha, \forall i, \\ & \mathbf{v}_i^T\mathbf{v}_i = 1, \ \forall i, \\ & \mathbf{v}_i \perp_{\bar{L}} \mathbf{v}_j, \forall i \neq j. \end{aligned} \tag{4}$$

α is the only parameter in this objective. It serves as a cut-off threshold. Any cut $\mathbf{v}$ that fails to satisfy (in the relaxed sense) at least α constraints in $\bar{Q}$ will be rejected. For the remaining cuts, which are called *feasible cuts* [18], the top-K ones with the lowest cost on $\bar{L}$ will be chosen. To guarantee the existence of at least K feasible cuts to choose from, when setting the value of α, we require:

$$\alpha \in [0, \lambda_K(\bar{Q})], \tag{5}$$

where $\lambda_K(\bar{Q})$ is the K-th largest eigenvalue of $\bar{Q}$. Note that since $\bar{Q}$ is PSD in our problem setting, if α is set to 0, any $\mathbf{v} \in \mathbb{R}^N$ is a feasible cut.

Given that Eq.(5) is satisfied, raising the threshold α will reduce the range of the feasible cuts that we can choose from. As a result, the final partition will be more biased towards the constraint matrix $\bar{Q}$. Similarly, lowering the threshold α will give the algorithm more freedom to choose cuts that are favored by the graph $\bar{L}$. Therefore, in practice, the choice of α is determined by our preference between $\bar{L}$ and $\bar{Q}$. If we have confidence that $\bar{Q}$ is more accurate, then we should set α to a larger value; and vice versa.

Next we present an alternative objective which is non-parametric. To get rid of the cut-off threshold α, we consider

the following cost-satisfaction ratio for any cut $\mathbf{v}$:

$$f(\mathbf{v}) \triangleq \frac{\mathbf{v}^T \bar{L} \mathbf{v}}{\mathbf{v}^T \bar{Q} \mathbf{v}}.$$

Since $\bar{Q}$ is now guaranteed to be PSD, we have

$$f(\mathbf{v}) \in [0, \infty), \forall \mathbf{v} \in \mathbb{R}^N.$$

The goal of constrained spectral clustering is to maximize $\mathbf{v}^T \bar{Q} \mathbf{v}$ and minimize $\mathbf{v}^T \bar{L} \mathbf{v}$, which is equivalent to minimize $f(\mathbf{v})$. Therefore $f(\mathbf{v})$ becomes a unified measure for the quality of the cut $\mathbf{v}$: smaller $f(\mathbf{v})$ means better cut.

Formally, the non-parametric version of our objective is:

$$\begin{aligned} \underset{\mathbf{v}_i \in \mathbb{R}^N | _{i=1}^K}{\text{argmin}} \quad & \sum_{i=1}^{K} \frac{\mathbf{v}_i^T \bar{L} \mathbf{v}_i}{\mathbf{v}_i^T \bar{Q} \mathbf{v}_i}, \\ \text{s.t.} \quad & \mathbf{v}_i^T \mathbf{v}_i = 1, \ \forall i, \\ & \mathbf{v}_i \perp_{\bar{L}} \mathbf{v}_j, \forall i \neq j. \end{aligned} \tag{6}$$

3.3 Efficient Solutions

We first show how to solve the non-parametric objective in Eq.(6). Consider the generalized eigenvalue problem:

$$\bar{L} \mathbf{v} = \lambda \bar{Q} \mathbf{v}. \tag{7}$$

Since both $\bar{L}$ and $\bar{Q}$ are Hermitian and PSD, we will have N real generalized eigenvectors [**?**], and they are the critical points for the generalized Rayleigh quotient [9]

$$\frac{\mathbf{v}^T \bar{L} \mathbf{v}}{\mathbf{v}^T \bar{Q} \mathbf{v}}. \tag{8}$$

Consequently, the solution to Eq.(6) is the top-K generalized eigenvectors from Eq.(7) that minimize the Rayleigh quotient in Eq.(8). The full algorithm is presented in Algorithm 2.

On the other hand, the solution to Eq.(4) is similar to that in [18]. We first solve the generalized eigenvalue problem

$$\bar{L} \mathbf{v} = \lambda (\bar{Q} - \alpha I) \mathbf{v},$$

and the non-negative eigenvectors are the feasible cuts. For efficiency consideration, in our implementation, we simply solve Eq.(7), and choose eigenvectors associated with eigenvalues that are no smaller than α. This allows us to only solve the eigenvalue problem once, and reuse the eigenvectors for multiple α values. The algorithm for the parametric objective in Eq.(4) is summarized in Algorithm 1.

3.4 An Illustrative Example

In Figure 1 we illustrate how our approach can improve the clustering on a single view. The dataset used are 1200 documents on 6 topics originally written in English and their translation into French. We plot the top-6 eigenvectors (each color corresponds to one eigenvector). The y-axis is the entries values of the eigenvectors. Due to the sparsity of the feature space, for spectral clustering on either single view, we can observe a couple of "spikes" among the top eigenvectors, which are trivial cuts that separate a small number of documents from the rest. Including these trivial cuts will inevitably harm the quality of the final partition. In contrast, when we combined two views together using our approach, the two views rejected the trivial cuts proposed by each other, and reached agreement on 6 eigenvectors that are much more informative. As a result, the quality of the final partition is substantially improved.

Note that Figure 1 is just one among many cases where our approach could help improve the clustering. It could be still effective even when there are no trivial cuts involved.

3.5 Remarks

3.5.1 Algorithm Complexity

The runtime for both algorithms are dominated by that of solving the eigenvalue problem in Eq.(7). Therefore, the complexity of our algorithm is on par with spectral clustering in big-O notation, which is $O(KN^2)$ for dense matrices and $O(KM)$ for sparse matrices, M be the number of non-zeros entries.

3.5.2 Direction of Knowledge Transfer

The process of enforcing the constraint matrix $\bar{Q}$ to the graph $\bar{L}$ can be viewed as the transfer of knowledge. From Eq.(4) we can see that the transfer is asymmetric. Therefore, given the original view and the translated view, we need to decide which view should be used to construct $\bar{Q}$ and which view should be used to construct $\bar{L}$. The role of the constraint matrix $\bar{Q}$ is to select N' feasible cuts from the N generalized eigenvectors ($N' \ll N$); the role of the graph Laplacian is to select K min-cuts from the N' feasible cuts. Therefore, $\bar{Q}$ plays a more critical role than $\bar{L}$ does. If the quality of $\bar{Q}$ is very poor, it will rule out eigenvectors that lead to good cuts; and once those cuts are ruled out, $\bar{L}$ will not be able to recover them. On the other hand, if the quality of $\bar{Q}$ is very high, it will select a set of good feasible cuts for $\bar{L}$ to choose from.

Consequently, in practice we should use the better view to construct the constraint matrix. Although such prior knowledge is not always available, according to our observation on real data, the original view usually has better quality. Therefore in our experiments, $\bar{Q}$ is always constructed from the original documents and $\bar{L}$ is always constructed from the translation.

4. EMPIRICAL STUDY

In this section, we empirically study the performance of our algorithm on real-world data and compare it to existing techniques. We aim to answer the following questions:

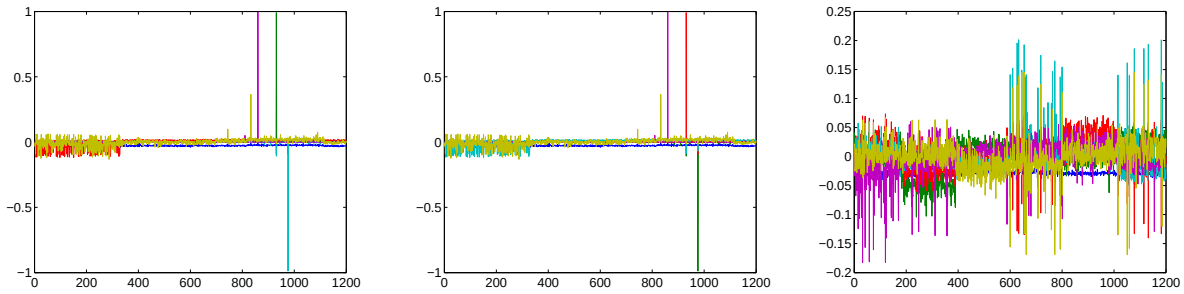
1. **Effectiveness:** Is our algorithm able to improve the clustering quality by using machine translation?
2. **Consistency:** Is the performance gain of our algorithm consistent over a range of diverse data samples?
3. **Comparison:** Does our algorithm outperform existing techniques?

4.1 Methodology

4.1.1 Dataset

We used the Reuters RCV1/RCV2 Multilingual dataset[1] introduced by [2]. This dataset has been used by previous work [2, 11, 13] to evaluate the performance of multiview spectral clustering algorithm. The dataset contains documents originally written in five different languages, namely

[1] `http://multilingreuters.iit.nrc.ca/ReutersMultiLingualMultiView.htm`



(a) Original (b) Translation (c) Original+Translation

Figure 1: An illustrative example to show how our algorithm can utilize the extra view to achieve better clustering. In each plot we show the top-6 eigenvectors from spectral clustering. The "spikes" in (a) and (b) are trivial cuts suggested by the two views individuals, which could cause a bad final partition. After combining the two views using our approach, these trivial cuts are ruled out in (c), which in turn leads to a better partition.

Table 2: Statistics of the dataset

Language	#docs	#words
English	18,758	21,531
French	26,648	24,839
German	29,953	34,279
Spanish	12,342	11,547
Italian	24,039	15,506

Topics	#docs	Percentage
C15	18,816	16.84
CCAT	21,426	19.17
E21	13,701	12.26
ECAT	19,198	17.18
GCAT	19,178	17.16
M11	19,421	17.39

English (`EN`), French (`FR`), German (`GR`), Spanish (`SP`) and Italian (`IT`). Each document, originally written in one language, was translated to the other four languages using the Portage system [16]. The documents are categorized into six different topics. The statistics of the dataset is summarized in Table 2. More detail can be found on the dataset homepage.

The dataset is provided in the form of tf-idf vectors. We did not apply additional preprocessing to the data. No dimensionality reduction technique was applied. We used cosine similarity to construct the graphs.

4.1.2 Evaluation metrics

We used two commonly used metrics to measure the quality of clustering, namely Adjusted Rand Index (ARI) [10] and Normalized Mutual Information (NMI). Both of them indicate the similarity between a given clustering and the ground truth partition: higher value means better clustering; 1 means perfect match.

4.1.3 Algorithm Implementations

We implemented both the parametric (`csp-p`) and the non-parametric (`csp-n`) version of our algorithm[2]. For the parametric version, we always set

$$\alpha \leftarrow \lambda_{2K}(\bar{Q}),$$

where $\lambda_{2K}(\bar{Q})$ is the $2K$-th largest eigenvalue of $\bar{Q}$. In other words, we provide $2K$ feasible cuts and $\bar{L}$ will choose the top-K with the lowest costs.

We also implemented in MATLAB five baseline algorithms to compare with:

- `orig`: Spectral clustering based on the original view only.
- `trans`: Spectral clustering based on the translated view only.
- `kersum`: The kernel summation algorithm for multi-view spectral clustering, which performs spectral clustering on the weighted sum of the two views' kernels. Previous study [6] showed that this approach works very well in practice, even in comparison to much more sophisticated multiview learning techniques. We used equal weights in our experiments.
- `mrw`: The mixing random walk algorithm proposed in [19], which finds the stationary distribution of a mixing random walk in both graphs. We used equal weights for the two views in our experiments.
- `co-reg`: The co-regularization multiview spectral clustering algorithm proposed in [13]. We implemented the centroid based version and used the centroids to compute final clustering. This algorithm has one parameter, which is the weight for the regularizer. We set it to 0.01 in our experiments, as suggested in the original paper.

4.1.4 Task Description

We first pick a language pair, say `EN-FR`, which means documents that are originally written in English along with their French translation. To maximize the diversity between the data samples we use, in each trial we randomly sample 1200 documents, which is less than 10% of all available documents. We have 100 trials for each language pair. We apply our algorithm and the baseline algorithms to the sample and

[2]The MATLAB code we used will be made available upon publication.

partition it into $K = 6$ clusters. We measure the resultant clusterings using both ARI and NMI. Since the last step of spectral clustering involves the K-means algorithm, in each trial, we repeat K-means algorithm 100 times with 100 random seeds and report the average performance.

For all language pairs, we report the average performance (Figure 2), aggregated results over 100 trials (Figure 3), and the performance of each individual trial (Figure 4).

4.2 Results and Analysis

Since we have 5 different languages, there are 20 possible original-translation language combinations. We show our results on 8 pairs, namely English to French (EN-FR), German (EN-GR), Italian (EN-IT), Spanish (EN-SP), and German to English (GR-EN), French (GR-FR), Italian (GR-IT), and Spanish (GR-SP). The conclusions we draw from these 8 pairs also hold for the other 12 pairs.

First we give an overview of the results in terms of average performance. In Figure 2, we report the average ARI of 7 different algorithms on 8 different language pairs. Our approach (`csp-p`) shows consistent and significant (99% confidence level) improvement over the clustering on the original view only (`orig`) for all language pairs. Also, for all language pairs, `csp-p` has the highest average ARI.

More detailed results are reported in Figure 3, with boxplots for all 7 algorithms and 8 language pairs, in terms of ARI and NMI, respectively. Besides of showing the advantage of our approach (`csp-p`), as we have seen in Figure 2, Figure 3 illustrates the diversity of the data samples we used in different trials. Some data samples were easier to cluster and others more difficult. This demonstrates that the effectiveness of our approach is not limited to a certain data distribution or a certain language.

Note that the performance of our non-parametric approach (`csp-n`) is not as good as the parametric one (`csp-p`). This is expected because `csp-n` uses zero prior knowledge. On the other hand, as shown in Figure 2, `csp-n` managed to outperform `orig` on all 8 language pairs, as well as the multiview competitors on several language pairs. This result is non-trivial considering the approach is completely parameter-free.

To further demonstrate the consistency and reliability of our approach over different random samples, in Figure 4, we show the trial by trial breakdown of the performance gain of our approach (`csp-p` and `csp-n`) over the clustering on the original view only (`orig`), as measured by ARI. We can see that for 800 random trials over 8 different language pairs, our algorithm achieved positive gain in most cases, and it rarely caused large performance loss. This means our approach is reliable in practice. Practitioners can apply our approach to a dataset, with peace of mind that it is very likely that our algorithm will improve the result, and it is very unlikely that our algorithm will lead to a great performance loss.

5. RELATED WORK

Recent work in the multiview learning literature studied the potential of using automated machine translation to improve both document classification [1, 2] and clustering [11–13]. Given a set of documents, their translations in another language is modeled as a second view. These multiple views are considered as partial observations of the same set of data objects. When properly combined, these views will complement each other and improve the resultant classification or clustering. Several multiview learning algorithms have been proposed and tested on the same Reuters RCV1/RCV2 Multilingual Dataset [2] as we used in this paper. Empirical results confirmed the helpfulness of machine translation. Note that the effectiveness of multiview learning is based on the assumption that the views are compatible and conditionally independent [5]. There is no practical way to validate either assumption for a specific sample.

In this work, we adopted an alternative approach, namely constrained clustering [3]. In contrast to the multiview formulation, constrained clustering does not make assumption about the underlying distribution of the data. Traditional constrained clustering algorithm cannot be directly applied to our problem because they can only deal with sparse and accurate constraints. If the number of constraints increases, or incorrect constraints are introduced, the clustering algorithm will be over-constrained [7]. It is also difficult to choose a small set of helpful constraints [8].

Wang and Davidson [18] proposed a spectral formulation for constrained clustering, which suits our problem setting well because it can incorporate soft constraints. Instead of enforcing each and every constraint given, they use a user-specified parameter to lower bound the number of satisfied constraints. As a result, noisy and incorrect constraints can be ignored by the final clustering. The difference between our work and theirs is that they considered sparse constraints generated from ground truth or domain experts, whereas our constraints are dense, generated from a distance metric. Furthermore, the constraint matrix we use is always PSD. As a result, we are able to extend their objective to K-way partition, and develop a new non-parametric solution to the problem.

Note that the focus of this work is to show that 1) machine translation is indeed helpful for document clustering and 2) how to use machine translation to improve document clustering. We assume that the translations of documents are readily available. Technical detail of statistical machine translation [14] is not considered, although it is expected that better translation will lead to greater performance gain.

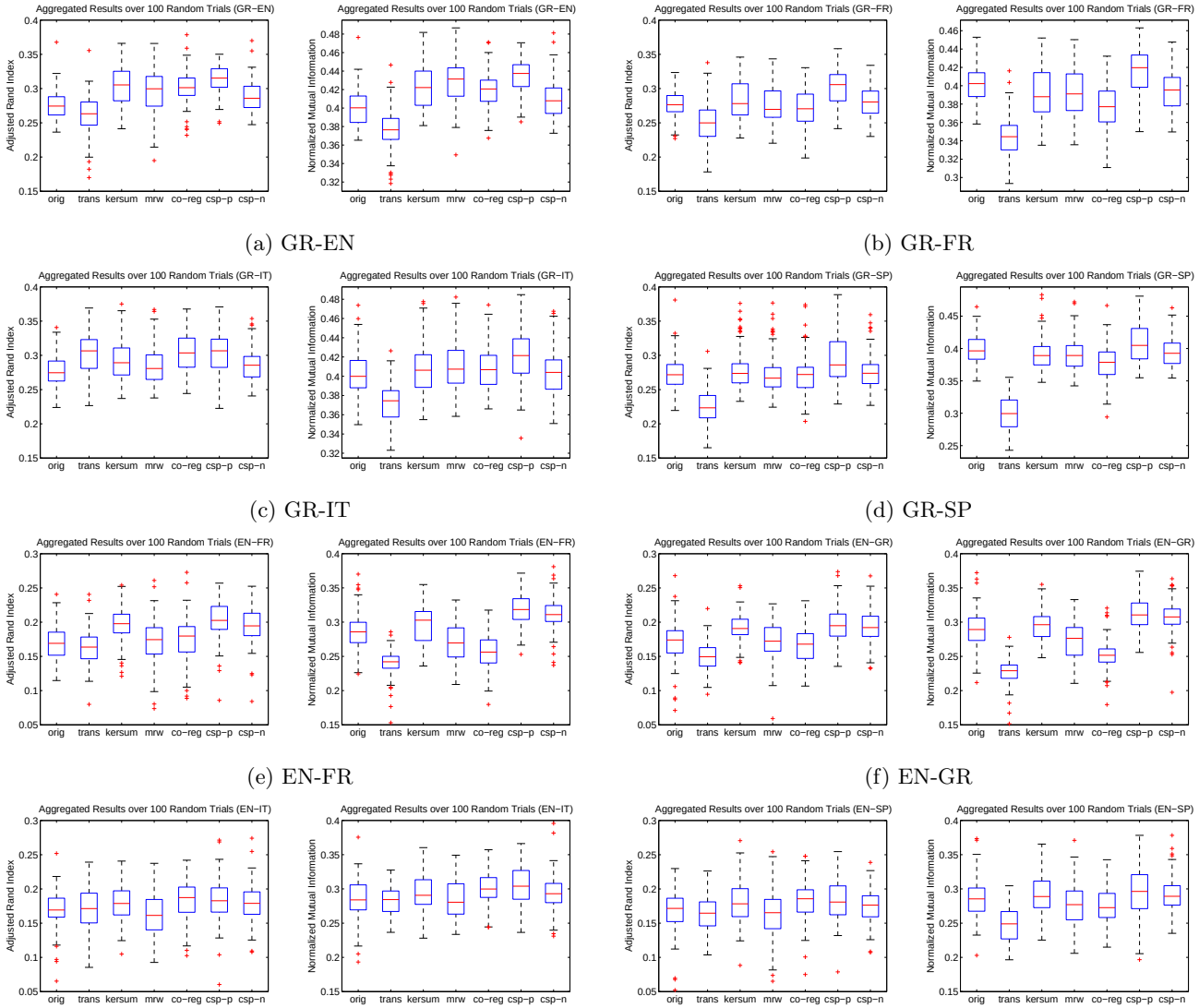
6. CONCLUSION AND FUTURE WORK

Automated machine translation offers the ability to supplement existing document representations with additional information. Previous work has explored using this additional information in a multiview clustering setting with some success. In this work, we take an alternative approach of encoding the additional information as constraints. This is a challenging problem since existing constrained clustering algorithms expect a small number of constraints generated from the ground truth or domain experts, whereas MT produces dense and potentially inaccurate information. We proposed two algorithms that can be viewed as an extension of spectral clustering to encode many noisy constraints without being over-constrained and with the ability to ignore constraints. We showed with real data that our approach is effective (it improves the clustering by just using the original documents, see Figure 2), consistent (since it can ignore poor side information, see Figure 3 and 4), and outperforms other comparable techniques (see Figure 2).

It is important to remember that existing work on multiview clustering [12] showed that the performance gain from using more than two views is marginal. Our future work will revisit this question by clustering using constraints gen-

	orig	trans	kersum	mrw	co-reg	csp-p	csp-n
GR-EN	0.276	0.261	0.303	0.296	0.303	**0.314**	0.289
GR-FR	0.277	0.250	0.284	0.279	0.274	**0.303**	0.281
GR-IT	0.279	0.301	0.294	0.286	0.303	**0.304**	0.287
GR-SP	0.274	0.224	0.280	0.273	0.271	**0.293**	0.276
EN-FR	0.168	0.163	0.195	0.172	0.174	**0.203**	0.194
EN-GR	0.171	0.149	0.192	0.172	0.167	**0.197**	0.192
EN-IT	0.168	0.170	0.178	0.161	0.183	**0.184**	0.179
EN-SP	0.169	0.163	0.180	0.164	**0.182**	**0.182**	0.175

Figure 2: A summary of the average ARI on 8 language pairs. Our approach (csp-p) consistently and significantly outperforms the clustering on the original view (orig) at 99% confidence level. It also outperforms its competitors (significantly in most cases). Detailed results are reported in Figure 3.



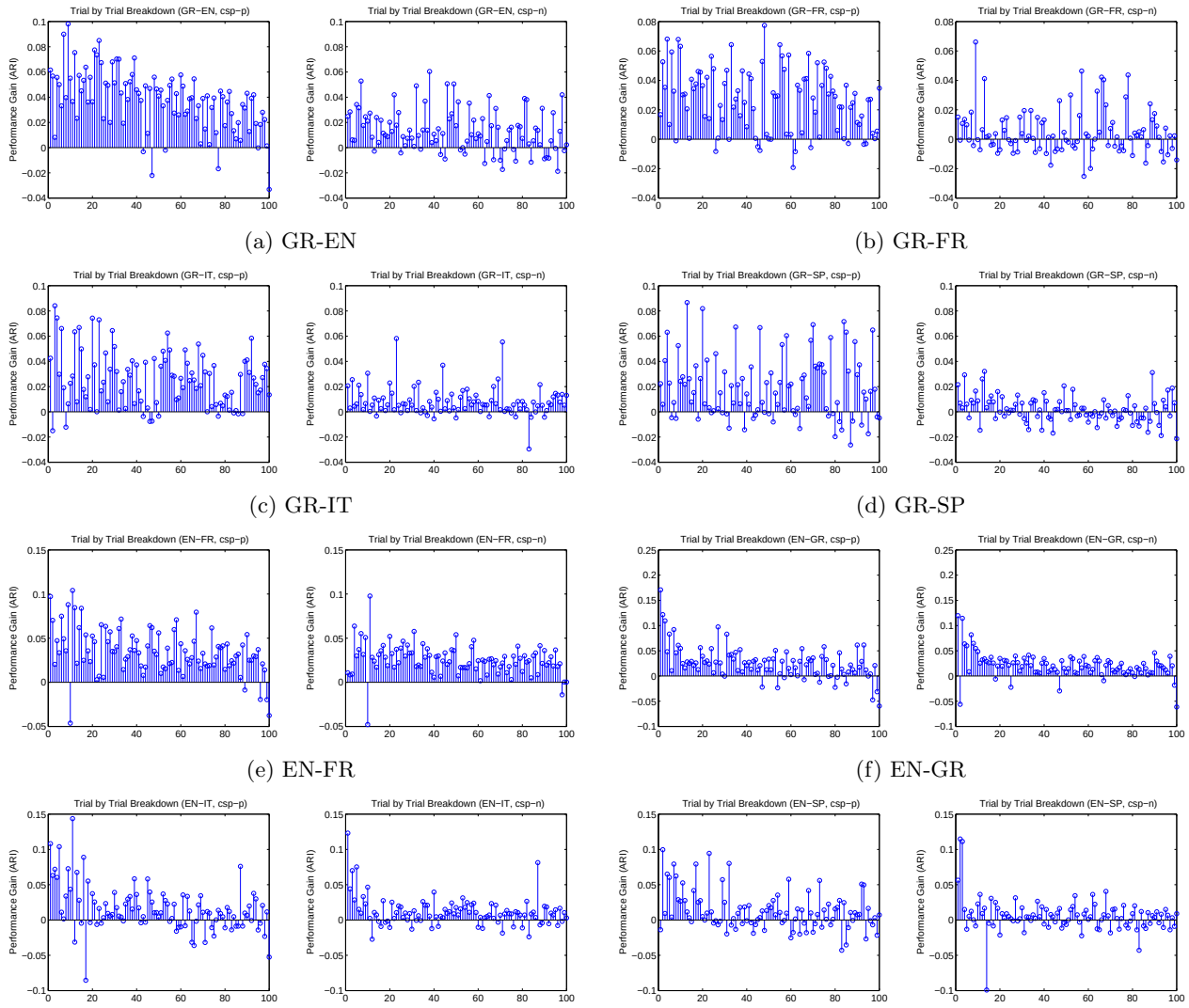
(a) GR-EN (b) GR-FR

(c) GR-IT (d) GR-SP

(e) EN-FR (f) EN-GR

(g) EN-IT (h) EN-SP

Figure 3: The box plot for 8 language pairs, 100 random trials each. Results are evaluated in terms of both ARI and NMI. Our technique (csp-p) consistently and significantly improves over the clustering on the original view (orig), and outperforms the competitors.



(a) GR-EN (b) GR-FR

(c) GR-IT (d) GR-SP

(e) EN-FR (f) EN-GR

(g) EN-IT (h) EN-SP

Figure 4: The trial by trial breakdown of the performance gain (w.r.t. `orig`) of our technique (`csp-p` and `csp-n`) on all 8 language pairs. Our technique not only achieved positive performance gain in average, but also in most of the individual trials.

erated from multiple translated views. Also, in this work we showed that adding constraints from translations does not hurt the clustering (since our approach ignores harmful constraints). However, an important problem we hope to address is determining *a priori* how much the translation will improve the clustering. This will help address the problem: "Which language should we translate into?"

7. ACKNOWLEDGMENTS

We gratefully acknowledge support of this research via ONR grants N00014-09-1-0712 Automated Discovery and Explanation of Event Behavior, N00014-11-1-0108 Guided Learning in Dynamic Environments and NSF Grant NSF IIS-0801528 Knowledge Enhanced Clustering.

8. REFERENCES

[1] M.-R. Amini and C. Goutte. A co-classification approach to learning from multilingual corpora. *Machine Learning*, 79(1-2):105–121, 2010.

[2] M.-R. Amini, N. Usunier, and C. Goutte. Learning from multiple partially observed views - an application to multilingual text categorization. In *NIPS*, pages 28–36, 2009.

[3] S. Basu, I. Davidson, and K. Wagstaff, editors. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*. Chapman & Hall/CRC, 2008.

[4] M. W. Berry, editor. *Survey of text mining: clustering, classification, and retrieval*. Springer, 2004.

[5] A. Blum and T. M. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, pages 92–100, 1998.

[6] C. Cortes, M. Mohri, and A. Rostamizadeh. Learning non-linear combinations of kernels. In *NIPS*, pages 396–404, 2009.

[7] I. Davidson and S. S. Ravi. Identifying and generating easy sets of constraints for clustering. In *AAAI*, pages 336–341, 2006.

[8] I. Davidson, K. Wagstaff, and S. Basu. Measuring constraint-set utility for partitional clustering algorithms. In *PKDD*, pages 115–126, 2006.

[9] R. Horn and C. Johnson. *Matrix analysis*. Cambridge Univ. Press, 1990.

[10] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[11] Y.-M. Kim, M.-R. Amini, C. Goutte, and P. Gallinari. Multi-view clustering of multilingual documents. In *SIGIR*, pages 821–822, 2010.

[12] A. Kumar and H. D. III. A co-training approach for multi-view spectral clustering. In *ICML*, pages 393–400, 2011.

[13] A. Kumar, P. Rai, and H. D. III. Co-regularized multi-view spectral clustering. In *NIPS*, pages 1413–1421, 2011.

[14] A. Lopez. Statistical machine translation. *ACM Comput. Surv.*, 40(3), 2008.

[15] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):888–905, 2000.

[16] N. Ueffing, M. Simard, S. Larkin, and J. H. Johnson. NRC's PORTAGE system for WMT 2007. In *In ACL-2007 Second Workshop on SMT*, pages 185–188, 2007.

[17] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[18] X. Wang and I. Davidson. Flexible constrained spectral clustering. In *KDD*, pages 563–572, 2010.

[19] D. Zhou and C. Burges. Spectral clustering and transductive learning with multiple views. In *ICML*, pages 1159–1166, 2007.