The NVI Clustering Evaluation Measure

Roi Reichart
ICNC
Hebrew University of Jerusalem
roiri@cs.huji.ac.il

Ari Rappoport
Institute of Computer Science
Hebrew University of Jerusalem
arir@cs.huji.ac.il

Abstract

Clustering is crucial for many NLP tasks and applications. However, evaluating the results of a clustering algorithm is hard. In this paper we focus on the evaluation setting in which a gold standard solution is available. We discuss two existing information theory based measures, V and VI, and show that they are both hard to use when comparing the performance of different algorithms and different datasets. The V measure favors solutions having a large number of clusters, while the range of scores given by VI depends on the size of the dataset. We present a new measure, NVI, which normalizes VI to address the latter problem. We demonstrate the superiority of NVI in a large experiment involving an important NLP application, grammar induction, using real corpus data in English, German and Chinese.

1 Introduction

Clustering is a major technique in machine learning and its application areas. It lies at the heart of unsupervised learning, which has great potential advantages over supervised learning. This is especially true for NLP, due to the high efforts and costs incurred by the human annotations required for training supervised algorithms. Recent NLP problems addressed by clustering include POS induction (Clark, 2003; Goldwater and Griffiths, 2007), word sense disambiguation (Shin and Choi, 2004), semantic role labeling (Baldewein et al., 2004), pitch accent type disambiguation (Levow, 2006) and grammar induction (Klein, 2005).

Evaluation of clustering results is a challenging task. In this paper we address the *external measures* setting, where a correct assignment of elements to *classes* is available and is used for evaluating the quality of another assignment of the elements into *clusters*. Many NLP works have used external clustering evaluation measures (see Section 2).

Recently, two measures have been proposed that avoid many of the weaknesses of previous measures and exhibit several attractive properties (see Sections 2 and 3): the VI measure (Meila, 2007) and the V measure (Rosenberg and Hirschberg, 2007). However, each of these has a serious drawback. The possible values of VI lie in [0, $2log\ N$], where N is the size of the clustered dataset. Hence it has limited use when comparing performance on different datasets. V measure values lie in $[0, 1]$ regardless of the dataset, but the measure strongly favors a clustering having many small clusters. In addition, V does not have many of the attractive properties of VI.

This paper has two contributions. First, we propose the NVI measure, a normalization of VI which guarantees that the score of clusterings that VI considers good lies in [0,1], regardless of dataset size. Most of VI's attractive properties are retained by NVI.

Second, we compare the behavior of V, VI and NVI in various situations to the desired behavior and to each other. In particular, we show that V gives high scores to clusterings with a large number of clusters even when they are of low quality. We demonstrate this both in a synthetic example (Section 5) and in the evaluation (in three languages) of a difficult NLP problem, labeled parse tree induc-

Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL), pages 165–173,
Boulder, Colorado, June 2009.

tion (Section 6). We show that in both cases, NVI constitutes a better clustering evaluation measure.

2 Previous Evaluation Measures

A large number of clustering quality measures have been proposed. Here we briefly survey the three main types, mapping based measures, counting pairs measures and information theory based measures.

We first review some terminology (Meila, 2007; Rosenberg and Hirschberg, 2007). In a **homogeneous** clustering, every cluster contains only elements from a single class. In a **complete** clustering, all elements of each class are assigned to the same cluster. The **perfect** solution is the fully homogeneous and complete clustering. We will illustrate the behavior of some measures using three extreme cases: the **single cluster** case, in which all data elements are put in the same single cluster; the **singletons** case, in which each data element is put in a cluster of its own; and the **no knowledge** case, in which the class distribution within each cluster is identical to the class distribution in the entire dataset. If the single cluster solution is not the perfect one, the no knowledge solution is the worst possible solution. Throughout the paper, the number of data elements to be clustered is denoted by **N**.

Mapping based measures are based on a post-processing step in which each cluster is mapped to a class. Among these are: L (Larsen, 1999), D (Van Dongen, 2000), misclassification index (MI) (Zeng et al., 2002), H (Meila, 2001), clustering F-measure (Fung et al., 2003) and micro-averaged precision and recall (Dhillon et al., 2003). As noted in (Rosenberg and Hirschberg, 2007), these measures evaluate not only the quality of the proposed clustering but also of the mapping scheme. Different mapping schemes can lead to different quality scores for the same clustering. Moreover, even when the mapping scheme is fixed, it can lead to not evaluating the entire membership of a cluster and not evaluating every cluster (Meila, 2007).

Counting pairs measures are based on a combinatorial approach which examines the number of pairs of data elements that are clustered similarly in the reference and proposed clustering. Among these are Rand Index (Rand, 1971), Adjusted Rand Index (Hubert and Arabie, 1985), Γ statistic (Hubert and Schultz, 1976), Jaccard (Milligan et al., 1983), Fowlkes-Mallows (Fowlkes and Mallows, 1983) and Mirkin (Mirkin, 1996).

Meila (2007) described a number of problems with such measures. The most acute one is that their values are unbounded, making it hard to interpret their results. The problem can be solved by transformations adjusting their values to lie in $[0, 1]$, but the adjusted measures suffer from severe distributional problems, again limiting their usability in practice.

Information-theoretic (IT) based measures are those addressed in this work. The measures in this family suffer neither from the problems associated with mappings, since they evaluate the entire membership of each cluster and not just a mapped portion, nor from the distributional problems of the counting pairs measures.

Zhao and Karypis (2001) define *Purity* and *Entropy* as follows:
$Purity = \sum_{r=1}^{k} \frac{1}{N} max_i(n_r^i)$
$Entropy = \sum_{r=1}^{k} \frac{n_r}{N}(-\frac{1}{logq}\sum_{i=1}^{q}\frac{n_r^i}{n_r}log(\frac{n_r^i}{n_r}))$
where q is the number of classes, k the number of clusters, n_r cluster r's size, and n_r^i is the number of elements in class i assigned to cluster r.

Both measures are good measures for homogeneity (Purity increases and Entropy decreases when homogeneity increases). However, they do not evaluate completeness at all. The singletons solution is thus considered optimal even if in fact it is of very low quality.

Dom (2001) proposed the Q measure, the sum of a homogeneity term $H(C|K)$ and a model cost term calculated using a coding theory argument:
$Q(C, K) = H(C|K) + \frac{1}{N}\sum_{k=1}^{|k|} log\binom{h(k)+|C|-1}{|C|-1}$
where C are the correct classes, K are the induced clusters and $h(k)$ is the number of elements in cluster k. Dom also presented a normalized version of the Q measure (called Q_2) whose range is $(0, 1]$ and gives higher scores to clusterings that are preferable. As noted by (Rosenberg and Hirschberg, 2007), the Q measure does not explicitly address the completeness of the suggested clustering. Due to the cost term, if two clusterings have the same $H(C|K)$ value, the model prefers the one with the lower number of clusters, but the trade-off between homogeneity and completeness is not explicitly addressed.

In the next section we describe the V and VI mea-

sures, which are IT measures that explicitly assess both the homogeneity and completeness of the clustering solution.

BCubed (Bagga and Baldwin, 1998) is an attractive measure that addresses both completeness and homogeneity. It does not explicitly use IT concepts and avoids mapping. In this paper we focus on V and VI; a detailed comparison with BCubed is out of our scope here and will be done in future work.

Several recent NLP papers used clustering techniques and evaluation measures. Examples include (Finkel and Manning, 2008), using VI, Rand index and clustering F-score for evaluating coreference resolution; (Headden et al., 2008), using VI, V, greedy 1-to-1 and many-to-1 mapping for evaluating unsupervised POS induction; (Walker and Ringger, 2008), using clustering F-score, the adjusted Rand index, V, VI and Q_2 for document clustering; and (Reichart and Rappoport, 2008), using greedy 1-to-1 and many-to-1 mappings for evaluating labeled parse tree induction.

Schulte im Walde (2003) used clustering to induce semantic verb classes and extensively discussed non-IT based clustering evaluation measures. Pfitzner et al. (2008) presented a comparison of clustering evaluation measures (IT based and others). While their analysis is extensive, their experiments were confined to artificial data. In this work, we experiment with a complex NLP application using large real datasets.

3 The V and VI Measures

The V (Rosenberg and Hirschberg, 2007) and VI (Meila, 2007) measures are IT based measures. In this section we give a detailed description of these measures and analyze their properties.

Notations. The partition of the N data elements into classes is denoted by $C = \{c_1, \ldots, c_{|C|}\}$. The clustering solution is denoted by $K = \{k_1, \ldots, k_{|K|}\}$. $A = \{a_{ij}\}$ is a $|C| \times |K|$ contingency matrix such that a_{ij} is the number of data elements that are members of class c_i and are assigned by the algorithm to cluster k_j.

As other IT measures, V and VI assume that the elements in the dataset are taken from a known distribution (both assume the uniform distribution), and thus the classes and clusters can be treated as random variables. When assuming the uniform distribution, the probability of an event (a class or a cluster) is its relative size, so $p(c) = \sum_{k=1}^{|K|} \frac{a_{ck}}{N}$ and $p(k) = \sum_{c=1}^{|C|} \frac{a_{ck}}{N}$. Under this assumption we can talk about the entropies $H(C)$ and $H(K)$ and the conditional entropies $H(C|K)$ and $H(K|C)$:

$$H(C) = -\sum_{c=1}^{|C|} \frac{\sum_{k=1}^{|K|} a_{ck}}{N} log \frac{\sum_{k=1}^{|K|} a_{ck}}{N}$$

$$H(K) = -\sum_{k=1}^{|K|} \frac{\sum_{c=1}^{|C|} a_{ck}}{N} log \frac{\sum_{c=1}^{|C|} a_{ck}}{N}$$

$$H(C|K) = -\sum_{k=1}^{|K|} \sum_{c=1}^{|C|} \frac{a_{ck}}{N} log \frac{a_{ck}}{\sum_{c=1}^{|C|} a_{ck}}$$

$$H(K|C) = -\sum_{k=1}^{|K|} \sum_{c=1}^{|C|} \frac{a_{ck}}{N} log \frac{a_{ck}}{\sum_{k=1}^{|K|} a_{ck}}$$

In Section 2 we defined the concepts of homogeneity and completeness. In order to satisfy the homogeneity criterion, each cluster must be contained in a certain class. This results in the minimization of the conditional entropy of the classes given the clusters, $H(C|K) = 0$. In the least homogeneous solution, the conditional entropy is maximized, and $H(C|K) = H(C)$. Similarly, in order to satisfy the completeness criterion, each class must be contained in a certain cluster, which results in the minimization of the conditional entropy of the clusters given the classes, $H(K|C) = 0$. In the least complete solution, the conditional entropy is maximized, and $H(K|C) = H(K)$.

The VI measure. Variation of information (VI) is defined as follows:

$$VI(C, K) = H(C|K) + H(K|C).$$

In the least homogeneous (complete) clustering, the values of $H(C|K)$ ($H(K|C)$) are maximal. As a clustering solution becomes more homogeneous (complete), the values of $H(C|K)$ ($H(K|C)$) decrease to zero. Consequently, *lower* VI values imply better clustering solutions. In the perfect solution, both $H(C|K) = 0$ and $H(K|C) = 0$ and thus $VI = 0$. For the least homogeneous and complete clustering solution, where knowing the cluster tells nothing about the class and vise versa, $VI = H(C) + H(K)$.

As a result, the range of values that VI takes is dataset dependent, and the numbers themselves tell

us nothing about the quality of the clustering solution (apart from a score of 0, which is given to the best possible solution).

A bound for VI values is a function of the maximum number of clusters in C or K, denoted by k^*. This is obtained when each cluster contains a single element, and $k^* = N$. Thus, $VI \in [0, 2logN]$. Consequently, the range of VI values is dataset dependent and unbounded when datasets change. This means that it is hard to use VI to compare the performance of a clustering algorithm across datasets.

An apparent simple solution to this problem would be to normalize VI by $2logk^*$ or $2logN$, so that its values would lie in $[0, 1]$. We discuss this at the end of the next section.

VI has two useful properties. First, it satisfies the **metric axioms**, that is: $VI(C, K) \geq 0$, $VI(C, K) = VI(K, C)$, $VI(C_1, C_2) + VI(C_2, C_3) \geq VI(C_1, C_3)$. This gives an intuitive understanding of the relation between VI values.

Second, it is **convexly additive**. This means that if K is obtained from C by splitting C_j into clusters $K_j^1, \ldots, K_j^m$, $\hat{H}(K_j) = -\sum_{i=1}^{m} P(K_j^i|C_j) logP(K_j^i|C_j)$, then $VI(C, K) = P(C_j)\hat{H}(K_j)$. This property guarantees that all changes to VI are local; the impact of splitting or merging clusters is limited only to those clusters involved, and its size is relative to the size of these clusters.

The V measure. The V measure uses homogeneity (h) and completeness (c) terms as follows:

$$h = \begin{cases} 1 & H(C) = 0 \\ 1 - \frac{H(C|K)}{H(C)} & H(C) \neq 0 \end{cases}$$

$$c = \begin{cases} 1 & H(K) = 0 \\ 1 - \frac{H(K|C)}{H(K)} & H(K) \neq 0 \end{cases}$$

$$V = \frac{2hc}{h + c}$$

In the least homogeneous clustering, $H(C|K)$ is maximal, at $H(C|K) = H(C)$. In this case h reaches its minimum value, which is 0. As homogeneity increases $H(C|K)$ values decrease. For the most homogeneous clustering, $H(C|K) = 0$ and $h = 1$. The same considerations hold for c, which ranges between 0 (for the least complete clustering) and 1 (for a complete clustering). Since V is defined to be the harmonic mean of h and c, V values lie in $[0, 1]$. Consequently, it can be used to compare the performance of clustering algorithms across datasets. Higher V values imply better clusterings.

Unlike VI, V does not satisfy the metric axioms and is not convexly additive. The range of values it can get does not depend on dataset size.

Extreme cases for the two measures. In the single cluster solution $H(C|K) = H(C)$ and $H(K|C) = 0$, and thus $V = 0$ (the worst possible score) and $VI = H(C)$. If there is indeed only a single class, then $VI = 0$, the best possible score, which is the correct behavior. VI behaves better than V here.

The singletons solution is a fully homogeneous clustering in which $H(C|K) = 0$. The score of each measure depends on the completeness of the solution. The completeness of a singletons clustering increases with the number of classes. In the extreme case where every element is assigned to a unique class ($|C| = |K| = N$) singletons is also complete, $H(K|C) = 0$, and $V(C, K) = 1, VI(C, K) = 0$. Both measures exhibit the correct behavior.

If there are classes that contain many elements, singletons is far from being complete and should be treated as a low quality solution. Again, in the singletons solution $VI = H(K|C)$. Suppose that the number of clusters is fixed. When the number of classes increases, this value decreases, which is what we want. When the number of classes decreases, the score increases, which is again the correct behavior. In Section 5 we show that this desired behavior shown by VI is not shown by V.

Both measures treat the no knowledge solution as the worst one possible: $V = 0$, and $VI = H(C) + H(K)$.

4 Normalized Variation of Information

In this section we define NVI, a normalization of VI. NVI is N-independent and its values for clusterings considered as good by VI lie in $[0, 1]$. Hence, NVI can be used to compare clustering performance across datasets. We show that NVI keeps the convex additivity property of VI but not its metric axioms.

Definition. We define NVI to be:

$$NVI(C,K) = \begin{cases} \frac{H(C|K)+H(K|C)}{H(C)} & H(C) \neq 0 \\ H(K) & H(C) = 0 \end{cases}$$

We define NVI to be $H(K)$ when $H(C) = 0$ to satisfy the requirements that NVI values decrease as C and K become more similar and that NVI would be 0 when they are identical[1].

Range and extreme cases. Like VI, NVI decreases as the clustering becomes more complete and more homogeneous. For the perfect solution, $NVI = 0$. In both the single cluster and the no knowledge solutions, $H(C|K) = H(C)$. Thus, in the former case $NVI = 1$, and in the latter $NVI = 1 + \frac{H(K)}{HC} \geq 1$.

For the singletons clustering case, $NVI = \frac{H(K|C)}{H(C)}$. Suppose that the number of clusters is fixed. When the number of classes increases, the numerator decreases and the denominator increases, and hence the score decreases. In other words, as the real solution gets closer to the singletons solution, the score decreases, which is the correct behavior. When the number of classes decreases, the score increases, which is again the correct behavior.

For any pair of clusterings K_1 and K_2, $VI(C,K_1) > VI(C,K_2)$ iff $NVI(C,K_1) > VI(C,K_2)$. This implies that only clustering solutions whose VI scores are better (i.e., numerically lower) than the score of the single cluster solution will be scored lower than 1 by NVI.

Note that NVI is meant to be used when there is a 'correct' reference solution. In this case $H(C)$ is constant, so the property above holds. In this sense, VI is more general, allowing us to compare any three clustering solutions even when we do not have a correct reference one.

To summarize:

1. All clusterings considered by VI to be of high quality (i.e., better than the single cluster solution) are scored by NVI in the range of $[0, 1]$.

2. All clusterings considered by VI to be of lower quality than the single cluster solution are scored higher than 1 by NVI.

3. The ordering of scores between solutions given by VI is preserved by NVI.

4. The behavior of NVI on the extreme cases is the desired one.

Useful properties. In Section 3 we saw that VI has two useful properties, satisfying the metric axioms and being convexly additive. NVI is not symmetric since the term in its denominator is $H(C)$, the entropy of the correct class assignment. Thus, it does not satisfy the metric axioms. Being convexly additive, however, is preserved. In the class splitting scenario (see convex additivity definition in Section 3) it holds that $NVI(C,K) = \frac{P(C_j)\hat{H}(K_j)}{H(C)}$. That is, like for VI, the impact of splitting or merging a cluster on NVI is limited only to those clusters involved, and its size is relative to the size of these clusters. Meila (2007) derived various interesting properties of VI from the convex additivity property. These properties generally hold for NVI as well.

$H(K)$ **normalization.** Normalizing by $H(C)$ takes into consideration the complexity of the correct clustering. Another normalization option would be to normalize by $H(K)$, which represents the induced clustering complexity. This normalization does not guarantee that the scores of the 'good' clusterings lie in a data-independent range.

Let us define NVIK(C,K) to be $\frac{VI(C,K)}{H(K)}$ if $H(K) > 0$ and $H(C)$ if $H(K) = 0$. Recall that in order for NVIK to be 0 iff C and K are identical, we must require that $NVIK = H(C)$ when $H(K) = 0$. In the no knowledge case, $NVIK = \frac{H(C)+H(K)}{H(K)} = \frac{H(C)}{H(K)} + 1 > 1$. In the single cluster solution, however, $NVIK = H(C)$ (since in this case $H(K) = 0$) which ranges in $[0, logN]$. This is a serious drawback of NVIK. In Section 6 we empirically show an additional drawback of NVIK.

$logN$ **normalization.** Another possible normalization of VI is by $2logN$ (or $2logk^*$), which is an upper bound on VI values. However, this results in the values of the measure being dependent on dataset size, so results on datasets with different sizes again cannot be compared. For example, take any C and K and split each element into two. All entropy values, and the quality of the solution, are preserved, but the scores given to the two K's (before and after

[1] $H(C) = 0$ iff C consists of a single class, and therefore $H(C) = H(K) = 0$ iff C (K) consists of a single class (cluster).

7	1	1	1	0	0	0	0	0	0
0	7	1	1	1	0	0	0	0	0
0	0	7	1	1	1	0	0	0	0
0	0	0	7	1	1	1	0	0	0
0	0	0	0	7	1	1	1	0	0
0	0	0	0	0	7	1	1	1	0
0	0	0	0	0	0	7	1	1	1
1	0	0	0	0	0	0	7	1	1
1	1	0	0	0	0	0	0	7	1
1	1	1	0	0	0	0	0	0	7

	V	VI	NVI	NVIK
Singletons	0.667	2.303	**1**	0.5
Solution R	0.587	1.88	**0.81**	0.81

Table 1: The clustering matrix of solution R (top), and the scores given to it and to the singletons solution by the four measures (bottom). Although solution R is superior, the score given by V to the singletons solution is much higher. NVI exhibits the most preferable behavior (recall that higher V values are better, as opposed to the other three measures).

the split) by such a normalized VI would be different. Since $H(C)$ is preserved, the scores given by NVI to the two K's are identical.

5 Problematic V Behavior Example

In this section we provide a synthetic example that demonstrates an undesireable behavior of V (and NVIK) not manifested by VI and NVI. Specifically, V favors solutions with a large number of clusters, giving them higher scores than to solutions that are evidently superior. In addition, the score given to the singletons solution is high in absolute terms.

To present the example, we use the matrix representation A of a clustering solution defined in Section 3. The entries in row i sum to the number of elements in class i, while those in column j sum to the number of elements in cluster j.

Suppose that we have 100 elements assigned to 10 classes such that there are 10 elements in each class. We consider two clustering solutions: the singletons solution, and solution R whose matrix is shown in Table 1 (top). Like the real solution, solution R also has 10 clusters each having 10 elements. Solution R is not very far from the correct solution, since each cluster has 7 elements of the same class, and the three other elements in a cluster are taken from a different class each and can be viewed as 'noise'. Solution R is thus much better than the singletons solution. In order not to rely on our own opinion, we have performed a simple human judgment experiment with 30 subjects (university graduates in different fields), all of whom preferred solution R[2].

The scores given by V, VI, NVI and NVIK to the two solutions are shown in Table 1 (bottom). V scores solution R as being worse than the singletons solution, and gives the latter a number that's relatively high in absolute terms (0.667). VI exhibits qualitatively correct behavior, but the numbers it uses are hard to interpret since they are N-dependent. NVI scores solution R as being better than singletons, and its score is less than 1, indicating that it might be a good solution.

6 Grammar Induction Experiment

In this section we analyse the behavior of V, VI, NVI and NVIK using a highly non-trivial NLP application with large real datasets, the unsupervised labeled parse tree induction (LTI) algorithm of (Reichart and Rappoport, 2008). We focus on the labeling that the algorithm finds for parsing constituents, which is a clustering of constituents.

Summary of result. We show that V gives about the same score to a labeling that uses thousands of labels and to labelings in which the number of labels (dozens) is identical or smaller than the number of labels in the reference evaluation set (an annotated corpus). Contrary to V, both NVI and VI give much better scores to the solutions having a smaller number of labels.

It could be argued that the total number of 'real' labels in the data is indeed large (e.g., because every verb exhibits its own syntactic patterns) and that a small number of labels is just an arbitrary decision of the corpus annotators. However, most linguistic theories agree that there is a prototypical level of generalization that uses concepts such as Noun Phrase and Verb Phrase, a level which consists of at most dozens of labels and is strongly manifested by real language data. Under these accepted assumptions, the scoring behavior of V is unreasonable.

[2]We must rely on people's expectations, since the whole point in this area is that clustering quality cannot be formalized in an objective, application-independent way.

	MDL+SC (T labels)					MDL+SC (P labels)					MDL labels				
Corpus	L	$=1$	<10	$<10^2$	$\geq 10^2$	L	$=1$	<10	$<10^2$	$\geq 10^2$	L	$=1$	<10	$<10^2$	$\geq 10^2$
WSJ10	26	0	0	3	23	8	0	0	0	8	2916	2282	2774	2864	52
NEGRA10	22	0	2	12	10	6	0	0	1	5	1202	902	1114	1191	11
CTB10	24	1	4	11	13	9	1	2	4	5	1050	816	993	1044	6

Table 2: The number of elements (constituents) covered by the clusters (labels) produced by the MDL+SC (T or P labels) and MDL clusterings. L is the total number of labels. Shown are the number of clusters having one element, less than 10 elements, less than 100 elements, and more than 100 elements. It is evident that MDL induces a sparse clustering with many clusters that annotate very few constituents.

	V			VI			NVI			NVIK		
Corpus	MDL	T	P	MDL	T	P	MDL	T	P	MDL	T	P
WSJ10	0.4	0.44	0.41	3.83	2.32	1.9	2.21	1.34	**1.1**	0.81	0.86	1.2
NEGRA10	0.47	0.5	0.5	2.56	1.8	1.4	1.51	1.1	**0.83**	0.76	0.96	1.1
CTB10	0.42	0.42	0.45	3	2.22	1.85	1.72	1.26	**1.1**	0.87	1.1	1.25

Table 3: V, VI, NVI and NVIK values for MDL and MDL+SC with T or P labels. V gives the three clusterings very similar scores. NVIK prefers MDL labeling. NVI and VI both show the expected qualitative behavior, favoring MDL+SC clustering with P labels. The most preferable scores are those of NVI, whose numbers are also the easiest to interpret.

The experiment. The LTI algorithm has three stages: bracketing, initial labeling, and label clustering. Bracketing is done from raw text using the unsupervised incremental parser of (Seginer, 2007). Initial labeling is done using the BMM model (Borensztajn and Zuidema, 2007), which aims at minimizing the grammar description length (MDL). Finally, labels are clustered to a desired number of labels using the k-means algorithm with syntactic features extracted from the initially labeled trees. We refer to this stage as MDL+SC (for 'syntactic clustering'). Using a mapping-based evaluation with two different mapping functions, the LTI algorithm was shown to outperform previous work on unsupervised labeled parse tree induction.

The MDL clustering step induces several thousand labels for corpora of several tens of thousands of constituents. The role of the SC step is to generalize these labels using syntactic features. There are two versions of the SC step. In one, the number of clusters is identical to the number of labels in the gold standard annotation of the experimental corpus. This set of labels is called T (for target) labels. In the other SC version, the number of labels is the minimum number of labels required to annotate more than 95% of the constituents in the gold standard annotation of the corpus. This set of labels is called P (for prominent) labels. Since constituent labels follow the Zipfian distribution, P is much smaller than T.

In this paper we run the LTI algorithm and evaluate its labeling quality using V, VI, NVI and NVIK. We compare the quality of the clustering induced by the first clustering step alone (the MDL clustering) to the quality of the clustering induced by the full algorithm (i.e., first applying MDL and then clustering its output using the SC algorithm for T or P labels)[3].

We follow the experimental setup in (Reichart and Rappoport, 2008), running the algorithm on English, German and Chinese corpora: the WSJ Penn Treebank (English), the Negra corpus (Brants, 1997) (German), and version 5.0 of the Chinese Penn Treebank (Xue et al., 2002). In each corpus, we used the sentences of length at most 10,[4] numbering 7422 (WSJ10), 7542 (NEGRA10) and 4626 (CTB10).

The characteristics of the induced clusterings are shown in Table 2[5]. The table demonstrates the fact that MDL labeling, while perhaps capturing the

[3]Note that our evaluation here has nothing to do with the evaluation done in (Reichart and Rappoport, 2008), which provided a comparison of the full grammar induction results between different algorithms, using mapping-based measures. We evaluate the labeling stages alone.

[4]Excluding punctuation and null elements, according to the scheme of (Klein, 2005).

[5]The number of MDL labels in the table differs from their numbers, since we report the number of unique MDL labels used for annotating correct constituents in the parser's output, while they report the number of unique labels used for annotating *all* constituents in the parser's output.

salient level of generalization of the data in its leading clusters, is extremely noisy. For WSJ10, for example, 2282 of the 2916 unique labels annotate only one constituent, and 2774 labels label less than 10 constituents. These 2774 labels annotate 14.4% of compared constituents, and the 2864 labels that annotate less than 100 constituents each, cover 30.7% of the compared constituents (these percentages are not shown in the table). In other words, MDL is not a solution in which almost all of the mass is concentrated in the few leading clusters; its tail occupies a large percentage of its mass.

MDL patterns for NEGRA10 and CTB10 are very similar. For MDL+SC with T or P labels, most of the induced labels annotate 100 constituents or more. We thus expect MDL+SC to provide better clustering than MDL; a good clustering evaluation measure should reflect this expectation.

Table 3 shows V, VI, NVI and NVIK scores for MDL and MDL+SC (with T or P labels). For all three corpora, V values are almost identical for the MDL and the MDL+SC schemes. This is in contrast to VI and NVI values that strongly prefer the MDL+SC clusterings, fitting our expectations (recall that for these measures, the lower the score, the better the clustering). Moreover, VI and NVI prefer MDL+SC with P labels, which again accords with our expectations, since P labels were defined as those that are more salient in the data (see above).

The patterns of NVI and VI are identical, since $NVI = \frac{VI}{H(C)}$ and $H(C)$ is independent of the induced clustering. However, the numbers given by NVI are easier to interpret than those given by VI. The latter are basically meaningless, conveying nothing about clustering quality. The former are quite close to 1, telling us that clustering quality is not that good but not horrible either. This makes sense, because the overall quality of the labeling induction algorithm is indeed not that high: using one-to-one mapping (the more forgiving mapping), the accuracy of the labels induced by MDL+SC is only 45–72% (Reichart and Rappoport, 2008).

NVIK, the normalization of VI with $H(K)$, is worse even than V. This measure (which also gives lower scores to better clusterings) prefers the MDL over MDL+SC labels. This is a further justification of our decision to define NVI by normalizing VI by $H(C)$ rather than by $H(K)$.

Corpus	$H(C)$	$H(K)$		
		MDL	T	P
WSJ10	1.73	4.72	2.7	1.58
NEGRA10	1.69	3.36	1.87	1.29
CTB10	1.76	3.45	2.1	1.48

Table 4: Class ($H(C)$) and cluster ($H(K)$) entropy for MDL and MDL+SC with T or P labels. $H(C)$ is cluster independent. $H(K)$ increases with the number of clusters.

Table 4 shows the $H(C)$ and $H(K)$ values in the experiment. While $H(C)$ is independent of the induced clustering and is thus constant for a given annotated corpus, $H(K)$ monotonically increases with the number of induced clusters. Since both NVIK and the completeness term of V are normalized by $H(K)$, these measures prefer clusterings with a large number of clusters even when many of these clusters provide useless information.

7 Conclusion

Unsupervised clustering evaluation is important for various NLP tasks and applications. Recently, the importance of the completeness and homogeneity as evaluation criteria for such clusterings has been recognized. In this paper we addressed the two measures that address these criteria: VI (Meila, 2007) and V (Rosenberg and Hirschberg, 2007).

While VI has many useful properties, the range of values it can take is dataset dependent, which makes it unsuitable for comparing clusterings of different datasets. This imposes a serious restriction on the measure usage. We presented NVI, a normalized version of VI, which does not have this restriction and still retains some of its useful properties.

Using experiments with both synthetic data and a complex NLP application, we showed that the V measure prefers clusterings with many clusters even when these are clearly of low quality. VI and NVI do not exhibit such behavior, and the numbers given by NVI are easier to interpret than those given by VI.

In future work we intend to explore more of the properties of NVI and use it in other real NLP applications.

References

Amit Bagga and Breck Baldwin, 1998. Entity-based cross-document coreferencing using the vector space

model. *ACL 98*.

Ulrike Baldewein, Katrin Erk, Sebastian Pado, and Detlef Prescher 2004. Semantic role labeling with similarity based generalization using EM–based clustering. *Senseval '04*.

Thorsten Brants, 1997. The NEGRA export format. *CLAUS Report, Saarland University*.

Gideon Borensztajn and Willem Zuidema, 2007. Bayesian model merging for unsupervised constituent labeling and grammar induction. Technical Report, ILLC. http: //staff.science.uva.nl/∼gideon/

Alexander Clark, 2003. Combining distributional and morphological information for part of speech induction. *EACL '03*.

I. S. Dhillon, S. Mallela, and D. S. Modha, 2003. Information theoretic co-clustering. *KDD '03*.

Byron E. Dom, 2001. An information-theoretic external cluster validity measure . *Journal of American statistical Association*,78:553–569.

Jenny Rose Finkel and Christopher D. Manning, 2008. Enforcing transitivity in coreference resolution. *ACL '08*.

E.B Fowlkes and C.L. Mallows, 1983. A method for comparing two hierarchical clusterings. *Journal of American statistical Association*,78:553–569.

Benjamin C. M. Fung, Ke Wang, and Martin Ester, 2003. Hierarchical document clustering using frequent itemsets. *SIAM International Conference on Data Mining '03*.

Sharon Goldwater and Thomas L. Griffiths, 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. *ACL '07*.

William P. Headden, David McClosky, and Eugene Charniak, 2008. Evaluating unsupervised part-of-speech tagging for grammar induction. *COLING '08*.

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

L. Hubert and J. Schultz, 1976. Quadratic assignment as a general data analysis strategy. *British Journal of Mathematical and Statistical Psychology*, 29:190–241.

Dan Klein, 2005. The unsupervised learning of natural language structure. Ph.D. thesis, Stanford University.

Bjornar Larsen and Chinatsu Aone, 1999. Fast and effective text mining using linear-time document clustering. *KDD '99*.

Gina-Anne Levow, 2006. Unsupervised and semi-supervised learning of tone and pitch accent. *HLT-NAACL '06*.

Marina Meila and David Heckerman, 2001. An experimental comparison of model-based clustering methods. *Machine Learning*, 42(1/2):9-29.

Marina Meila, 2007. Comparing clustering – an information based distance. *Journal of Multivariate Analysis*, 98:873–895.

C.W Milligan, S.C Soon and L.M Sokol, 1983. The effect of cluster size, dimensionality and the number of clusters on recovery of true cluster structure. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 5:40–47.

Boris G. Mirkin, 1996. Mathematical classification and clustering. *Kluwer Academic Press.*

Darius M. Pfitzner, Richard E. Leibbrandt and David M.W Powers, 2008. Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems: An International Journal*, DOI 10.1007/s10115-008-0150-6.

William Rand, 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statstical Association*, 66(336):846–850.

Roi Reichart and Ari Rappoport, 2008. Unsupervised induction of labeled parse trees by clustering with syntactic features. *COLING '08*.

Andrew Rosenberg and Julia Hirschberg, 2007. V–Measure: a conditional entropy–based external cluster evaluation measure. *EMNLP '07*.

Sabine Schulte im Walde, 2003. Experiments on the automatic induction of German semantic verb classes. *Ph.D. thesis, Universitat Stuttgart*.

Yoav Seginer, 2007. Fast unsupervised incremental parsing. *ACL 07*.

Sa-Im Shin and Key-Sun Choi, 2004. Automatic word sense clustering using collocation for sense adaptation. *The Second Global WordNet Conference*.

Stijn van Dongen, 2000. Performance criteria for graph clustering and markov cluster experiments. *Technical report CWI, Amsterdam*

Daniel D. Walker and Eric K. Ringger, 2008. Model-based document clustering with a collapsed Gibbs sampler. *KDD '08*.

Nianwen Xue, Fu-Dong Chiou and Martha Palmer, 2002. Building a large–scale annotated Chinese corpus. *ACL '02*.

Yujing Zeng, Jianshan Tang, Javier Garcia-Frias, and Guang R. Gao, 2002. An adaptive meta-clustering approach: combining the information from different clustering results. *IEEE Computer Society Bioinformatics Conference (CSB '02)* .

Ying Zhao and George Karypis, 2001. Criterion functions for document clustering: experiments and analysis. *Technical Report TR 01-40, Department of Computer Science, University of Minnesota.*