Fuzzy V-Measure – An Evaluation Method for Cluster Analyses of Ambiguous Data

Jason Utt, Sylvia Springorum, Maximilian Köper, Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung
Universität Stuttgart
[uttjn|riestesa|koepermn|schulte]@ims.uni-stuttgart.de

Abstract

This paper discusses an extension of the V-measure (Rosenberg and Hirschberg, 2007), an entropy-based cluster evaluation metric. While the original work focused on evaluating hard clusterings, we introduce the Fuzzy V-measure which can be used on data that is inherently ambiguous. We perform multiple analyses varying the sizes and ambiguity rates and show that while entropy-based measures in general tend to suffer when ambiguity increases, a measure with desirable properties can be derived from these in a straightforward manner.

Keywords: clustering, evaluation, ambiguous data

1. Motivation

Ambiguity is ubiquitous in language and thus methods for dealing with ambiguous data are essential for robust systems and accurate representations in natural language processing. Many well-developed machine learning methods employ *clustering* as a main or pre-processing step. While simple clustering methods are often directly applied to unambiguous data (e.g. in computer vision), ambiguous data poses a problem in that in general, multiple labels will apply to a single data point. One common workaround is to simply assume the data can be represented unambiguously and to assign simple labels, e.g. in the case of verb classification (Merlo and Stevenson, 2001; Schulte im Walde, 2006). On our view, soft clustering techniques represent the most natural strategy for representing ambiguous data.

An important issue for such an investigation using soft clustering approaches, is the necessity of suitable evaluations of the soft cluster analyses, which are less developed and have not seen widespread acceptance so far. A standard evaluation method such as purity,

$$pur(C) = \frac{1}{N} \cdot \sum_j \max_k |c_j \cap g_k| \tag{1}$$

which computes the average overlap of a cluster with any group, does not intuitively fit a situation in which most class members belong to multiple classes.

Our work aims to fill this gap. We propose to extend the V-measure (Rosenberg and Hirschberg, 2007), which is an entropy-based measure developed for unlabeled cluster evaluation to handle ambiguous data, i.e. data which belongs to multiple classes in the gold standard.

The structure of the paper is as follows: Section 2 discusses the intuitions underlying the V-measure. Section 3 then expands on these intuitions defining the more general fuzzy V-measure. An analysis of the fuzzy V-measure is presented in Section 4. In Section 5, we further address the issues underlying all evaluations based on contingency tables and present a possible solution. Section 6 concludes the paper.

2. Entropy-based measure: V-Measure

Besides V-measure, there exist other information-based measures – e.g., *Variance of Information (VI)* (Meilă, 2007), with variants suggested by (Vinh et al., 2010) – however, we consider the intuitions and computational simplicity of V-measure, i.e. entropy gain and loss, to be useful.

In addition to the standard evaluation of a cluster analysis against a gold standard set of classes, the V-measure also allows for the comparison of two completely independent clusterings – with no restrictions in their similarity, the number of data points, or the number of clusters. In this paper we will adopt the terminology of clusters being compared against *classes*, assuming we have a gold standard classification of our data.

$v(C)$, the V-measure of a clustering C, of a set of data points is defined as a weighted mean of two complementary properties of the two partitionings of the data set. Each of these two constitute a particular desirable property for a clustering. The first is termed *homogeneity*,

$$hom(C) = \begin{cases} 1 & \text{if } H(C,G) = 0; \\ 1 - \frac{H(C|G)}{H(C,G)} & \text{else} \end{cases} \tag{2}$$

which gives a measure of how homogeneous the clusters in the clustering are. Here, H is the standard entropy: $H(C|G)$ denotes the conditional entropy of C given G and quantifies the amount of additional information contained in C with respect to G. The joint entropy, $H(C,G)$, is used for normalization. The second measure, *completeness* (cf. Equation 3), captures how intact the gold standard classes remain with respect to the clustering:

$$com(C) = \begin{cases} 1 & \text{if } H(G,C) = 0 \\ 1 - \frac{H(G|C)}{H(G,C)} & \text{else} \end{cases} \tag{3}$$

Homogeneity. In effect, homogeneity can be viewed as a generalization of the purity measure, which is a normalized measure (by the number of points N) to which degree each cluster c_j contains only members of one class g_k. For hom this corresponds to the amount of information the cluster contains about the class, which is high if the conditional entropy of the gold classes given the clustering, i.e. $H(G|C)$, is low. If each cluster contains only objects from one gold-standard class, then the entropy is at its minimum, $H(G|C) = 0$. This represents a maximally homogeneous clustering.

Completeness. Similar to the definition of homogeneity, completeness measures how well the classes map clusters within a cluster analysis. In the case where each gold-standard class maps only to one cluster, the clustering adds no additional information, $H(C|G)$ and is at its minimum. This represents a maximally complete clustering, in that each gold-standard class is completely covered by a particular cluster.[1] The final V-measure value is then computed as a weighted harmonic mean of the two homogeneity and completeness values:

$$v_\beta(C) = \frac{(1+\beta)\cdot hom(C)\cdot com(C)}{\beta\cdot hom(C) + com(C)} \tag{4}$$

In this paper, we give homogeneity and completeness equal weight ($\beta = 1$),

$$v(C) = \frac{2\cdot hom(C)\cdot com(C)}{hom(C) + com(C)} \tag{5}$$

but this can be freely chosen for a particular task depending on which measure is to be given priority.

It should be noted that in the final calculation step for hom and com, the polarity is reversed, i.e. when the respective conditional entropies are small, then the measure is at its maximum value 1, and 0 in the opposite case, that is, when there is no shared information and the conditional entropies equal the joint entropy.

In order to calculate these entropy values, we must define the joint and conditional probabilities across clusters and gold-standard classes. In (Rosenberg and Hirschberg, 2007), the joint probability of a cluster c and a gold-standard class g was estimated as

$$\hat{p}(c,g) = \frac{|c\cap g|}{N}, \tag{6}$$

where $|c\cap g|$ is the number of data points shared by c and g, and N is the total number of data points. This represents a problem in the case of ambiguous data, however, as there are more class memberships than data points. We will now illustrate this issue with an example.

3. Fuzzy V-Measure

Example with ambiguous data. Suppose we have a data set with four points: p_1, p_2, p_3, p_4. These points belong to four different gold-standard classes g_1, g_2, g_3, g_4 as shown in Figure 1. That is g_1 and g_4 each contain two members p_1, p_2 and p_2, p_4, respectively; g_2 has three members, namely p_1, p_3, p_4 and g_3 contains both p_2 and p_3.

Due to the ambiguity of our data, there are data points which belong to multiple classes, i.e. they are *fuzzy*. In order to calculate the probability as for the traditional V-measure, we would have to use a different normalizing constant, as the intersections of the different clusters are not disjoint, i.e. $\sum_{j,k}|c_j\cap g_k| > N$. At the same time, such an approach would give too much weight to highly ambiguous objects such as p_2. E.g. we would assign the same joint probability to the pair p_2 and g_4 as to p_4 and g_4. Obviously, this is unrealistic: p_4 belongs to only two classes while, p_2 belongs to three. We should thus give p_2 less weight as evidence for a particular class. Our approach is straightforward: we assign each point a total mass of 1 which is then evenly distributed among its classes, cf. Table 1. As this explicitly deals with fuzzy data, we term the resulting metric the *fuzzy V-measure*.

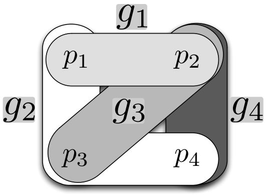


Figure 1: Distribution of ambiguous data.

	g_1	g_2	g_3	g_4
p_1	.5	.5	0	0
p_2	.33	0	.33	.33
p_3	0	.5	.5	0
p_4	0	.5	0	.5

Table 1: Distribution of data points in gold standard.

We thus generalize the counting of the original V-measure to a mass function μ:

$$\hat{p}(c,g) = \frac{\mu(c\cap g)}{M}, \tag{7}$$

where $\mu(c\cap g)$ is the total mass of the objects in the data shared by c and g, and M is the total mass of the clustering. Note that M will only be equal to N if each data point belongs to exactly as many classes as clusters. Using the cluster analysis shown in Figure 2, we can perform the calculation of V-measure on this data set. We see that cluster c_1 contains p_1, and p_2, and c_2 contains p_1, p_3 and p_4. Using the new gold standard mass distributions given in Table 2, we can build the contingency for clusters c_1, c_2 in Table 2. In this table, we see the masses for each intersection as explained above, i.e. cell i,j contains $\mu(c_i\cap g_j)$. This then serves to compute the joint and conditional probabilities.

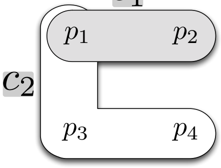


Figure 2: Clustering of ambiguous data.

[1] Note that Equations 2 and 3 differ from those in (Rosenberg and Hirschberg, 2007) in the denominators of the *else* condition because there were typos in the definitions (personal communication with Andrew Rosenberg).

	g_1	g_2	g_3	g_4	$\sum$
c_1	.83	.5	.33	.33	$= 2$
c_2	.5	1.5	.5	.5	$= 3$

Table 2: Contingency table containing mutual evidence between classes and clusters.

Now we can clearly see the advantage of our approach: While both c_1 and c_2 each share two points with the gold-standard classes g_1 and g_2 respectively, the higher ambiguity of p_2 in the first case means there is less evidence for c_1 given g_1 than c_2 given g_2, namely: $\hat{p}(c_1|g_1) = .83/2 < 1/2 = 1.5/3 = \hat{p}(c_2|g_2)$. Using these probabilities, we can easily compute the entropy values necessary for the calculation of the V-measure. This constitutes the fuzzy calculation of V-measure for a soft clustering. While the traditional V-measure .014 the fuzzy V score is .047. Both scores are small, but this is because all data points are ambiguous.

It should be noted that the fuzzy V-measure proposed here is applicable not only when data is ambiguous with respect to the gold standard classes themselves, but it also allows for the application to soft clusterings. We have already implemented and applied this measure (Springorum et al., 2013) to soft clusterings of highly ambiguous data, namely German prepositions. In such cases, the data points may be present in multiple clusters and simply add their respective mass to the cells in the contingency table.

The following section investigates the performance of fuzzy V when applied to clusterings and gold standards of varying sizes and across multiple rates of ambiguity.

4. Analysis of Fuzzy-V

The above example gives a general notion of distributing a data point's mass across classes in the contingency table resulting in a higher evaluation score. In this section, we apply the two V measures in different settings to test the stability of this result.

4.1. Experiment 1

As basis for our investigations, we build artificial data, by approximating the ambiguity rates as exhibited in actual linguistic data. The ambiguity rate in a data set is the distribution of class memberships – i.e. ambiguities – over all data points. In the example above, the ambiguity rate would be $2, 2, 2, 3$, as three points have an ambiguity of 2, and one point an ambiguity of 3. Figure 3 shows the ambiguity rates across parts of speech in WordNet 3.0 (Fellbaum, 1998). The automatically constructed ambiguity rates for our experiment were designed to give realistic ambiguity rates independent of data size, while being easy to construct and interpret: In the first step half of the data points are assigned an ambiguity rate of 1, i.e. they are unambiguous. Then, successively, the remaining data set is split in two halves, the first half is assigned an ambiguity rate one higher than the previous half. E.g., for 7 points, the automatically generated ambiguity rates would be 1, 1, 1, 1, 2, 2, 3. In this experiment, we keep the ambiguity rate of the data set constant, while varying its size.

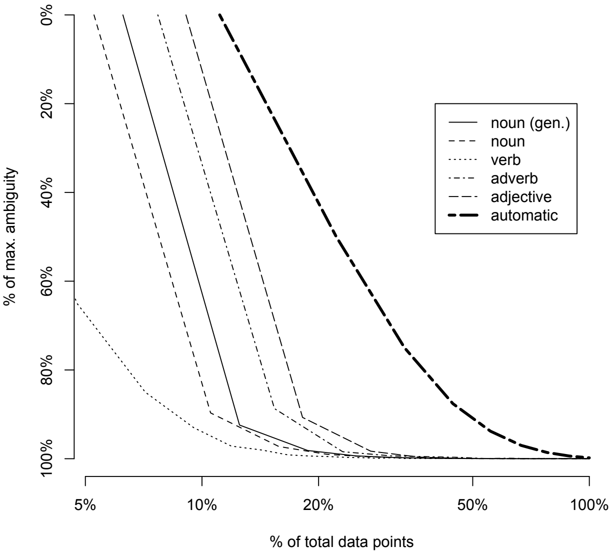
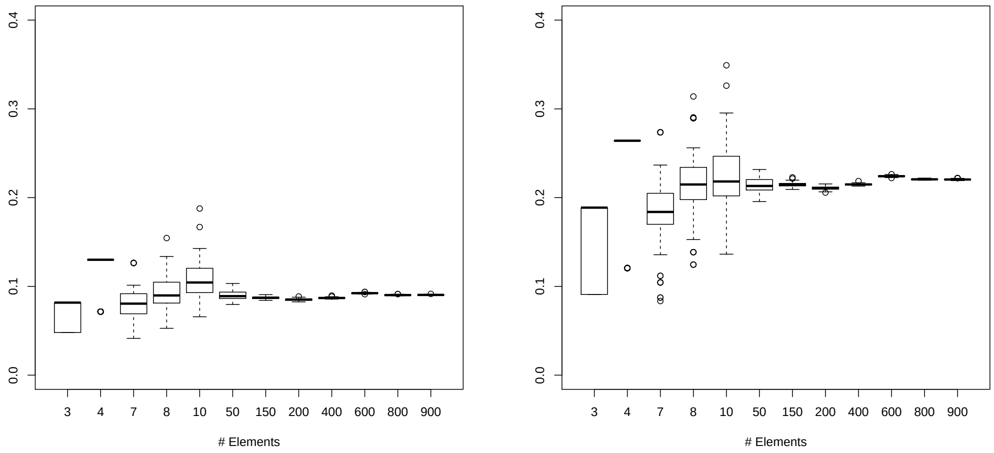


Figure 3: Ambiguity rates across parts of speech in WordNet 3.0. Data points are ordered left to right from less ambiguous to most ambiguous. Noun generalizations are taken from CoreLex (Buitelaar, 1998).

Data. We construct data sets with sizes ranging from 2 to approximately 1000 such that the ambiguity rates of the data points lie on the line in Figure 3. For each set of data points with their corresponding ambiguity rates, we randomly generate 100 gold standard classifications. These are selected uniformly from the $2^{|G|\times|C|}$ possible assignments for all classes (G) and clusters (C) for the given data points. We assume a *perfect clustering* for each data set, i.e. the clusters contain the same items as the classes in the gold standard. In such cases we would like to have scores close or equal to 1.

Evaluation. For each gold standard, together with its identical clustering, we evaluate the clustering using the traditional V measure as well as with our fuzzy V across the different assignments. As fuzzy V explicitly allows the assignment of multiple items to different classes (or clusters), we expect fuzzy V to reliably yield higher scores than traditional V.

Result. As can be seen in Figure 4, none of the measures reach the maximum of value 1, though the clusters are perfect in that they represent the same partitioning of the data set as the gold classes. However, our hypothesis is confirmed in such a way as that the values for fuzzy V are consistently higher than for V. The figures also show that with increasing data sizes, the variation for both fuzzy V and V decreases, while the overall variance of fuzzy V is a greater than for traditional V.



(a) Traditional V

(b) Fuzzy V

Figure 4: Impact of the number of datapoints

4.2. Experiment 2

While in the first experiment we assumed a perfect clustering, we now turn to arbitrary clusterings of the data.

Data. To show how well the measures capture the variation in matching the clusters to the classes, two clusterings with random object assignments are evaluated against each other, keeping the ambiguity rate constant, across the different data sizes. In total we compare 499 different assignments, starting from an assignment with only two elements and ending with an assignment containing 500 elements.

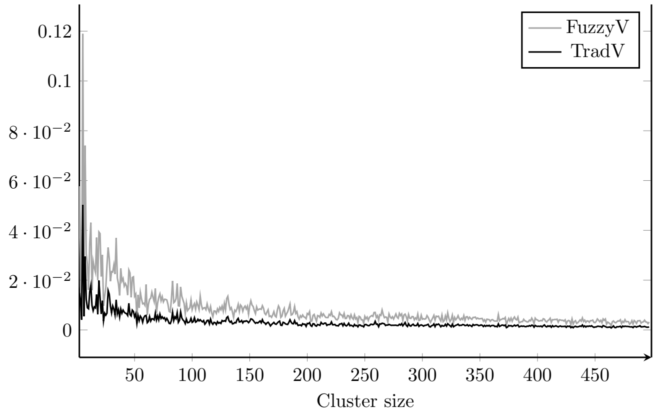


Figure 5: Experiments with unequal assignments

Results. Figure 5 shows that independent of how well the clustering maps to the gold classes, fuzzy V is less sensitive to the ambiguity than traditional V.[2]

4.3. Experiment 3

The goal of this experiment is to determine the effect of how ‘hard’ or ‘soft’ the clustering is on the resulting scores, i.e. how the number of ambiguous assignments impact both V measures. While the preceding experiment compared contrasting clusterings, while maintaining the data points’ ambiguity rates in the clustering, this experiment will vary the ambiguity rates in the clusters, i.e. the number of cluster assignments for each data point, among the test clusterings. The assignment of clusters to the gold standard classes is kept stable.

Setup. We begin with a hard clustered data set, where each data point is assigned to only one cluster and each cluster is correctly associated with one gold standard class. This cluster is evaluated against the gold standard containing many ambiguous elements. Then we incrementally add a new cluster assignment, according to the gold standard, until we get the ‘perfect’ soft clustering, i.e. identical with the gold standard. This experiment compares the behavior of the fuzzy V to the traditional V measure with a stepwise increase of the ambiguity rate. The gold standard for this purpose comprises 500 clusters for a total of 500 elements, where 250 elements are assigned to more than one gold standard class. In each step, the new, softer clustering is evaluated against the original gold standard. At each step, the assignments can be considered correct in the sense that they are assigned to one correct cluster (according to the gold standard); however, the spectrum of assignments each point’s polysemy would allow for, is captured incompletely.

Result. Figure 6 shows that with each clustering closer to the fuzzy gold standard, both values decrease. In the case of perfect clustering, again, both V measures yield smaller results compared to the previously less soft clustering. At first glance, this is a surprising outcome. However, it can readily be explained: As both measures are computed directly from entropies, the increased spread of mass in the contingency table due to ambiguity, leads to an increase in the overall uncertainty in the correspondence between clusters and classes.

If such soft clusterings are to be assigned a perfect score of

[2] It is important to note that there is no simple relationship between the two measures, e.g. by a constant factor, as tested in an additional experiment. Both heavily depend on the properties of the data, in particular the ambiguity rates, as will be shown in Experiment 3.

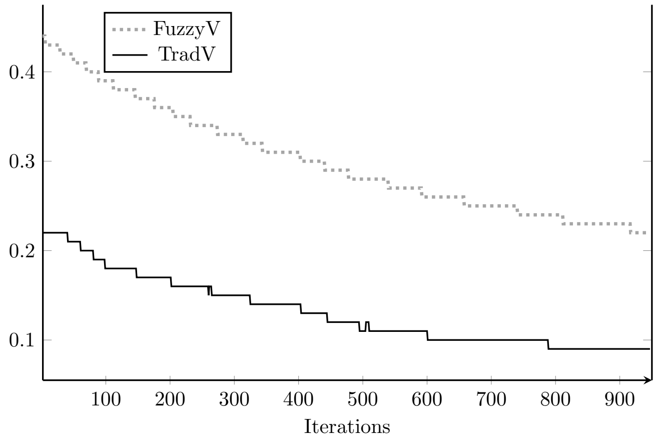


Figure 6: The impact of an increasing number of polysemous elements

1, the measures must be extended.In Section 5 we propose one possible extension of V measure which is able to solve this issue.

5. Beyond Entropy

As seen in the previous section, cluster evaluations carried out on partitions containing ambiguous elements are not assigned the desired maximum score of 1. So far, this is the case for both traditional V measure and the proposed fuzzy V. This section explains the underlying reasons for this behavior and introduces a method to extend evaluation metrics based on contingency tables, such as the fuzzy V and V measures, to mitigate this effect. The reason for this unfavorable behavior, is grounded in the construction of the contingency table. Such a traditional contingency table is in fact more consistent with a hard classification.

Assuming a perfect hard clustering where each column and each row of the contingency table yields only one value greater than zero, as in Figures 7a and 7b. In contrast, Figure 7c shows a contingency table in the case of a perfect clustering containing one ambiguous element.

$$\begin{array}{c c} & \begin{array}{ccc} g_1 & g_2 & g_3 \end{array} \\ \begin{array}{c} c_1 \\ c_2 \\ c_2 \end{array} & \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix} \end{array}$$

(a)

$$\begin{array}{c c} & \begin{array}{ccc} g_1 & g_2 & g_3 \end{array} \\ \begin{array}{c} c_1 \\ c_2 \\ c_2 \end{array} & \begin{pmatrix} 0 & 2 & 0 \\ 0 & 0 & 2 \\ 2 & 0 & 0 \end{pmatrix} \end{array}$$

(b)

$$\begin{array}{c c} & \begin{array}{ccc} g_1 & g_2 & g_3 \end{array} \\ \begin{array}{c} c_1 \\ c_2 \\ c_2 \end{array} & \begin{pmatrix} \boxed{1} & 2 & 0 \\ 2 & \boxed{1} & 0 \\ 0 & 0 & 2 \end{pmatrix} \end{array}$$

(c)

Figure 7: Example contingency tables

The gold classes g_1 and g_2 share one ambiguous element. This element leads to similarity between them and thus to double entries between several cluster/gold-class pairings (c_1:g_1,g_2 and c_2:g_1,g_2), which leads to a score less than 1. As seen previously, while fuzzy V is able to smooth this behavior it still does not provide the optimal score.

5.1. Dissimilarity

The previously explained problems are mainly caused by the way traditional contingency tables are constructed. This construction must, however, be extended to solve the issues encountered above, making the scoring based on such tables more reliable for soft clusterings. We introduce two additional steps:

1. Force a one-to-one mapping between cluster and gold-classes $c_x \rightarrow g_x$. This pairing should prefer combinations providing a high similarity and a low dissimilarity within the pair.

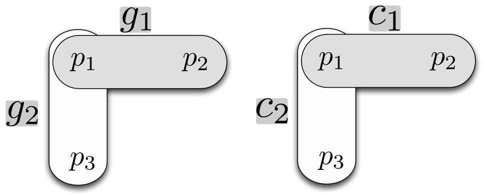
2. Penalize other mappings by uniformly distributing the remaining error mass (e_x), where e_x is defined as the dissimilarity between the best mapping c_x and g_x.

It is necessary to keep track of not only this *similarity* between classes and clusters, i.e., their shared elements' mass, but also of their *dissimilarity*, namely the missing and remaining elements between all cluster/class combinations. For any combination, a good mapping should lead to a high similarity and low dissimilarity. The difference between similarity and dissimilarity would then represent a more clear representation of the quality each cluster/class combination. Based on this score, the highest value determines the mapping between cluster and gold class. This information is enough to modify the contingency table. In a first step, we set all entries in all cells to zero, except for the best mapping entries. The cell containing the best mapping keeps the value of the traditional scoring scheme, i.e. its similarity. This step can be seen as removing unnecessary similarity, caused by ambiguous elements.

So far we have been optimistic because the contingency table includes only correct decisions, namely the score for the best mapping. To allow a complete cluster evaluation it is also necessary to punish wrong decisions. This is done by distributing error mass among the zero-entries in the contingency table. Error mass should include wrong decisions, such as missing elements or additional elements. This information is already captured in the previously calculated dissimilarity. Since the mapping for each cluster/class is already assigned, we only have to distribute the dissimilarity from that specific mapping. Note that this value is always zero in cases of perfect clustering.

Example. Consider the following example with three elements: p_1, p_2, p_3 where p_1 is ambiguous between two classes. The clustering corresponds exactly with the gold standard, as shown in Fig 8. The resulting contingency tables for both cases (hard and soft) are provided in Figures 8b and 8c.

The corresponding dissimilarity values are shown in Fig 9. The final mapping between clusters and gold classes is then simply via the highest score, where the score (Fig 10a and Fig 10b) is defined as the difference between similarity and



(a) Gold classes and clustering

$$sim = \begin{array}{c|cc} & g_1 & g_2 \\ c_1 & 2 & 1 \\ c_2 & 1 & 2 \end{array}$$

(b) 'Hard' contingency table

$$sim = \begin{array}{c|cc} & g_1 & g_2 \\ c_1 & 1.5 & 0.5 \\ c_2 & 0.5 & 1.5 \end{array}$$

(c) 'Soft' contingency table

Figure 8: Example: Clustering identical to gold classes

dissimilarity. In the example, this will lead to the mapping $c_1 \rightarrow g_1$ and $c_2 \rightarrow g_2$.

$$diss = \begin{array}{c|cc} & g_1 & g_2 \\ c_1 & 0 & 1 \\ c_2 & 1 & 0 \end{array}$$

(a) Hard dissimilarity matrix

$$diss = \begin{array}{c|cc} & g_1 & g_2 \\ c_1 & 0 & 0.5 \\ c_2 & 0.5 & 0 \end{array}$$

(b) Soft dissimilarity matrix

Figure 9: Dissimilarity tables for hard and soft clustering

The error mass is defined as the dissimilarity value for the best mapping. In this case this value is 0. In cases where the error mass is greater than 0, the error mass is distributed equally among the non-zero entries in each row. As seen in experiments 4.1., this addition can be used to extend both fuzzy V and traditional V measure.

$$score = \begin{array}{c|cc} & g_1 & g_2 \\ c_1 & \boxed{2} & 0 \\ c_2 & 0 & \boxed{2} \end{array}$$

(a) hard score

$$score = \begin{array}{c|cc} & g_1 & g_2 \\ c_1 & \boxed{1.5} & 0 \\ c_2 & 0. & \boxed{1.5} \end{array}$$

(b) soft score

Figure 10: The scores determine the mapping

Figure 11 shows the performance of the dissimilarity enhancement applied to the same experimental setup as in Figure 6 of Section 4.3. We can now see that the measures both converge toward the desired score of 1 for perfect soft clusterings.

6. Conclusion

As expected, fuzzy V consistently yields higher scores than traditional V in cases where it is to be anticipated. However, while it captures and correctly treats the ambiguous nature of the tested data, fuzzy V still suffers from the issues which emerge from calculating entropies from a standard contingency table.

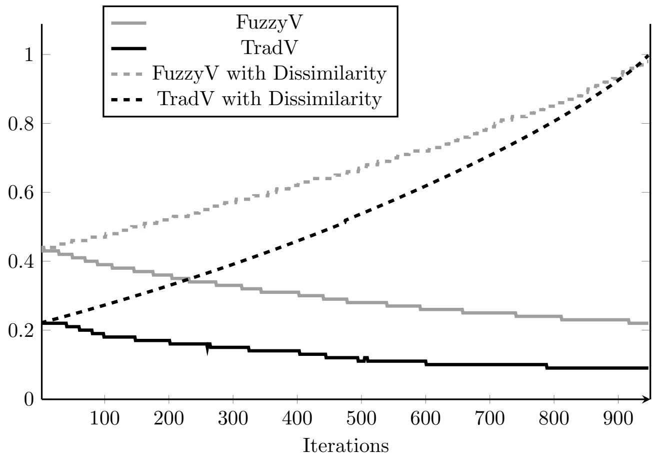


Figure 11: The impact of an increasing number of polysemous elements and the usage of dissimilarity

These findings illustrate the limits of such a purely on entropy-based measure. The complexity of the general uncertainty of ambiguous objects cannot be captured using only these methods. Thus, the addition of a further disambiguation step – namely on the level of class/cluster assignment – is required to better assess the quality of the clustering. This is similar to the calculation of purity, which uses the maximum intersection size between a cluster and the classes, which can be viewed as an implicit class-cluster assignment.

We have proposed a natural extension of the entropy-based V measure, the fuzzy V measure, which can handle better the evaluation of soft clusterings of ambiguous data. In addition, we highlighted the inherent drawbacks of entropy-based evaluation metrics of ambiguous classifications and have shown that these can be further improved upon using dissimilarity tables. As unlabeled ambiguous data is pervasive in NLP (e.g. in semantic classification, topic labeling), we feel this is a valuable addition to the evaluation techniques in this field.

7. Acknowledgements

The work has been supported by the DFG, Jason Utt and Sylvia Springorum were supported by SFB 732 and Maximilian Köper and Sabine Schulte im Walde were supported by DFG Heisenberg Fellowship SCHU-2580/1-1.

8. References

Paul Buitelaar. 1998. CoreLex: An Ontology of Systematic Polysemous Classes. In *Proceedings of the First International Conference on Formal Ontology in Information Systems (FOIS '98)*, Amsterdam. IOS Press.

Christiane Fellbaum, editor. 1998. *Wordnet: An Electronic Lexical Database*. Bradford Books.

Marina Meilă. 2007. Comparing Clusterings – an Information Based Distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Paola Merlo and Suzanne Stevenson. 2001. Automatic Verb Classification Based on Statistical Distributions of Argument Structure. *Computational Linguistics*, 27(3):373–408.

Andrew Rosenberg and Julia Hirschberg. 2007. V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420.

Sabine Schulte im Walde. 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. *Computational Linguistics*, 32(2):159–194.

Sylvia Springorum, Sabine Schulte im Walde, and Jason Utt. 2013. Detecting Polysemy in Hard and Soft Cluster Analyses of German Preposition Vector Spaces. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, Nagoya, Japan.

Nguyen Xuan Vinh, Juien Epps, and James Bailey. 2010. Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance. *Journal of Machine Learning Research*, 11:2837–2854.