A new Metric for the Evaluation of Dialog Act Classification*

Stephan Lesch and **Thomas Kleinbauer** and **Jan Alexandersson**
DFKI GmbH
Stuhlsatzenhausweg 3, D-66123 Saarbücken
{janal,kleiba,lesch}@dfki.de

Abstract

The standard evaluation metrics for dialog act classifiers are based on the boolean outcome of the exact classification. For multidimensional tag sets, such as the ICSI-MRDA tag set, this is stricter than necessary, since the misclassification might be partial and this can be good enough for the application in which the classifier is embedded. We propose a new forgiving metric and show some preliminary results. Some future work is sketched.

1 Introduction

We are concerned with the evaluation of automatic classification of utterances for multidimensional tag sets. Contrary to one-dimensional tag sets, such as the one developed within the VerbMobil project (Alexandersson et al., 1998), multidimensional tag sets assign not only one tag per utterance segment but a combination of a general tag and zero or more additional tags. This is the case for the ICSI meeting recorder dialog act tag set (henceforth MRDA), see (Shriberg et al., 2004).

When faced with a real-life application using speech, the task of assigning the correct tags can be further complicated through the absense of sentence boundaries. In addition to the dialog act labeling, the classifier might have to determine the segment boundaries, too, that constitute each utterance to be labeled (see (Ang et al., 2005)). Evaluation of such a task therefore needs to consider both the segmentation performance and the tagging results.

*The research presented here is funded by by the EU under the grant FP6-506811 (AMI).

For the pre-segmented case, the performance of the tagger is usually measured with *precision*, *recall*, e. g., (Reithinger and Klesen, 1997), and sometimes their harmonic mean, *fScore*. All three metrics are based on a notion of a "correct" classification which usually means that the tagger returned the correct label. This makes evaluation a binary function—the tagger output is either correct or incorrect.

For multidimensional tag sets the case is a bit more complex: each dimension in a label should be evaluated independently. For example, if the correct label is $\{t_1, t_2, t_3\}$ and the tagger assigns $\{t_3, t_4\}$, then dimension $\{t_3\}$ was classified correctly, dimensions $\{t_1, t_2\}$ were missed and $\{t_4\}$ was hallucinated. To compute the above measures within such a tag set, the size of the intersection between the assigned label and the actual label is divided by the size of the classified set in case of precision and the size of the correct set for recall ($\frac{1}{2}$ and $\frac{1}{3}$ in the above example). The *fScore* is still the harmonic mean between these two metrics (here $\frac{2}{5}$).

If we investigate the behaviour of the *fScore* metric, we see that whereas the value of a correctly assigned label is 1, and a completely erroneously assigned label is 0. Partly correct labels receive a different value depending on the size of the set of tags in the true tag. This is caused by the asymmetric behaviour of precision and recall. To highlight this, we use a small artifical tag set consisting of a general tag, T, and a set of additional tags $\{t_1, t_2, \ldots, t_6\}$ (see figure 1).

Table 1 shows the values for two fixed instances of the true label (first column). In the first case, the truth is $\{T, t_1\}$—written Tt_1— and in the second we have Tt_1t_3. The second row shows possible tagger output, alongside the *precision*, *recall* and *fScore* values for each result. We can observe an

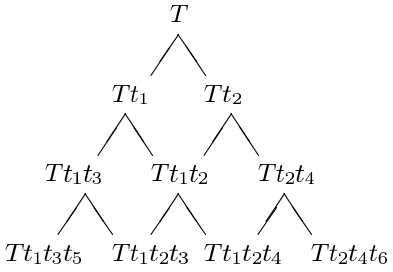

Figure 1: An excerpt of a made-up multidimensional tag set viewed as a lattice spanned by the subset relation. T is the general tag, and t_n are additional tags.

asymmetric behavior of *fScore* in rows 3 and 7. In both cases, the classified label contains one hallucinated special tag compared to the true label, but the *fScore* metric delivers different scores (0.8 and 0.86). A similar effect can be seen in rows 1 and 5, where in both cases the classified label misses one dimension in comparison to the ground truth while *fScore* yields values of 0.67 and 0.8.

Truth	Classified	*Prec*	*Rec*	*fScore*
Tt_1	T	1	0.5	0.67
Tt_1	Tt_1	1	1	1
Tt_1	Tt_1t_2	0.67	1	0.8
Tt_1t_3	T	1	0.33	0.5
Tt_1t_3	Tt_1	1	0.67	0.8
Tt_1t_3	Tt_1t_3	1	1	1
Tt_1t_3	$Tt_1t_2t_3$	0.75	1	0.86

Table 1: Values for *precision*, *recall* and *fScore* with different truth tags.

These effects occur because *fScore* takes the length of the true label into account (see also section 3): not only the absolute number of erroneously classified tags is relevant, but also the number of those that were classified *correctly*. In our example, row 3 yields two correct tags while row 7 has three—under this view, a higher *fScore* value in row 7 is justified. But it's also legitimate to ask for an evaluation metric that treats a deviation of one tag between classified label and truth equally, independent of

- whether the classified label contains one tag *too much* or *too little*.
- the length of the truth label, i.e. the position of this label in the hierarchy.

The rest of the paper is concerned with a new symmetric metric—SCORRE— which addresses the above points. We compare the behavior of our new metric based on experiments on the ICSI meeting corpus. The paper is organized as follows: Section 2 discusses the hierarchical view of tag sets. We recapitulate the standard metrics precision, recall and *fScore* in section 3. Section 4 is devoted to our new metric. Before we conclude the paper and point at future directions, we present an experiment and compare our results in section 5.

2 Multidimensional Tag Set Hierarchies

Our MRDA taggers for the ICSI meeting corpus currently obtain around 50% correct classifications (i.e. the label produced by the tagger is identical to the human annotation). An examination of the result reveals that another 30% of the classifications are very similar to the human annotations.

Multidimensional labels can be regarded as sets of tags, and it is thus possible to compare them by looking at their intersection and the differences between them. Likewise, the labels can be organized into a hierarchy similar to figure 1. There, the number of edges between two labels, ancestor relations, in particular, whether two nodes have a common ancestor, play a crusial role. For a hierarchy on multidimensional labels defined by the subset relation between labels, there is an obvious equivalence to the set comparison.

In our approach, we use lattices as a more general structure to express other relations between tags not based on subset, and still use distances to measure similarity between labels.

In case of the MRDA tagset, there are labels which we regard as incompatible although they share some aspects. For instance, if the general tag is erroneously tagged, we want to consider the classification entirely wrong, even if the true and the classifier label share some additional tags.

Also, a metric based on distances can as well be used on one-dimensional labels which are ordered in a hierarchy. This is the case for the Verbmobil labels, which fall into several groups, such as, suggestions, feedbacks, informs, or politeness. Also, these group labels do not have to be actual DA labels, but can be introduced for the sole purpose of comparing more specific labels.

3 Classifier Evaluation

The performance of a classifier is usually measured with respect to two orthogonal aspects: the overall performance on a test corpus and the performance per tag. For both aspects, the common measures

recall, *precision* and *fScore* can be used. For the *per-tag* performance, three values have to be computed:
• *tagged*(*label*)—the number of times the label was assigned by the classifier,
• *occurs*(*label*)—the number of times the label occurs in the test corpus, and
• *correct*(*label*)—the number of times the label was correctly assigned by the classifier.

$$\begin{aligned} \mathit{Precision}(\mathit{label}) &:= \frac{\mathit{correct}(\mathit{label})}{\mathit{tagged}(\mathit{label})} \\ \mathit{Recall}(\mathit{label}) &:= \frac{\mathit{correct}(\mathit{label})}{\mathit{occurs}(\mathit{label})} \\ \mathit{fScore}(\mathit{label}) &:= \frac{2 * \mathit{Prec}(\mathit{label}) * \mathit{Recall}(\mathit{label})}{\mathit{Prec}(\mathit{label}) + \mathit{Recall}(\mathit{label})} \end{aligned}$$

To evaluate a classifier's overall performance on a test corpus, it is necessary to compute the overlap between the classified label (DA^C) and ground truth (DA^T) for each segment. In the case of multidimensional dialog acts, we regard each label as a set of tags, and thus define the intersection $DA^I := DA^T \cap DA^C$. Similar to the *per-label* case, *precision* and *recall* measure the amount of missed and hallucinated tags.

$$\mathit{Precision}(DA^T, DA^C) := \frac{|DA^I|}{|DA^C|} \quad (1)$$

$$\mathit{Recall}(DA^T, DA^C) := \frac{|DA^I|}{|DA^T|} \quad (2)$$

Next, we base our definition on the distance in the hierarchy and rewrite (1) and (2) using the subset relation: Let

$$\begin{aligned} \delta^C &:= |DA^C| - |DA^I| \\ \delta^T &:= |DA^T| - |DA^I| \end{aligned}$$

then

$$\mathit{Precision}(DA^T, DA^C) = 1 - \frac{\delta^C}{|DA^C|} \quad (3)$$

$$\mathit{Recall}(DA^T, DA^C) = 1 - \frac{\delta^T}{|DA^T|} \quad (4)$$

$$\begin{aligned} \mathit{fScore}(DA^T, DA^C) &= \\ & \frac{2 * \mathit{Prec}(DA^T, DA^C) * \mathit{Rec}(DA^T, DA^C)}{\mathit{Prec}(DA^T, DA^C) + \mathit{Rec}(DA^T, DA^C)} \quad (5) \\ &= \ldots \\ &= 1 - \frac{\delta^C + \delta^T}{|DA^C| + |DA^T|} \quad (6) \end{aligned}$$

Here, the reason for the asymmetrical behaviour of *recall*, *precision* and *fScore* is obvious: the denominators relate the distances to the total complexity of the labels, that is, the fraction of the total information missed by the classifier and how much information not present in the truth was hallucinated by the classifier respectively.

(3), (4) and (6) show that we can view *recall*, *precision* and *fScore* as distance metrics: tags missing in the classified label—δ^T—reduces *recall*, while tags hallucinated by the classifier—δ^C—reduces *precision*. *fScore* is a mixture of both distances.

4 A Hierarchy-Based Distance Metric

In a lattice of labels in which each pair of labels (DA^C, DA^T) has a least upper bound DA^{lub}, we define δ^T and δ^C using the shortest paths between the labels and DA^{lub}:

$$\begin{aligned} \delta^C &:= |\mathit{minpath}(DA^C, DA^{lub})| \\ \delta^T &:= |\mathit{minpath}(DA^T, DA^{lub})| \end{aligned}$$

For a lattice defined by the subset relation between tags (Y is a child of X iff Y contains all tags in X, and exactly one additional tag), DA^{lub} is equivalent to the intersection DA^I and the set-differences are equivalent to the distances between DA^T/ $DA^C$and DA^I.

We now define a metric with a constant denominator:

$$\textsc{Scorre}(DA^T, DA^C) := 1 - \frac{\delta^C + \delta^T}{2 * \mathit{depth}}$$

if DA^{lub} exists, 0 otherwise. The denominator is a constant, i.e., normalization is done with the distance between two labels into the range between 1 (DA^C=DA^T) and 0 (maximum distance between DA^C and DA^T, or no path at all).

Note, that *depth* must be large enough to prevent the metric from going below zero. One possible choice is the maximum possible path length (e.g. the maximum number of possible tags in a label). However, this number may be large, and in practice, a smaller value may be as appropriate, as long as no longer distances occur in a classification experiment.

Finally, we define Scorracy of a classifier on a test corpus with n segments, true labels DA_i^T and classified labels DA_i^C:

$$\textsc{Scorracy} := \frac{\sum_{i=1}^{n} \textsc{Scorre}(DA_i^T, DA_i^C)}{n}$$

Thus, Scorracy is the mean distance between the DA_i^T and DA_i^C normalized to the range between 1 and 0.

5 An experiment

When building a statistical tagger for MRDA labels, we have to choose between two basic approaches—one is to treat the labels as monolithic units (i.e. the roughly 118000 utterances in the ICSI corpus are annotated with ca. 1250 different labels), while the other is to decompose the labels into the 55 different tags, build one classifier for each tag (or for a group of mutually exclusive tags), and compose the results from these classifiers into labels.

Preliminary experiments indicate that the monolithic tagger performs better in terms of correct classifications (ca. 3%). For the combined tagger, however, the sum of exact + partial matches is slightly better. SCORRACY indicates that the mean distance between truth and classifier guess is nearly the same for both classifiers. (Clark and Popescu-Belis, 2004) reports a similar experiment with an abstraction of these labels (the MALTUS tagset), with similar results: they obtain 73.2% correct classifications with a simplified variant of the MALTUS tagset, and only 70.5% with a combined classifier.

In our experiments, we have used $depth = 5$, since labels deeper in the hierarchy did not occur. The advantage in this choice is that SCORRE is easier to interpret intuitively that way; for instance, 0.8 means that the distance between DA^T and DA^C is 2.

	monolithic		combined
	MALTUS	MRDA	MRDA
correct	67.1%	51.4%	48.5%
underspec.	11.2%	19.8%	25.8%
overspecific	2.7%	3.2%	2.9%
neighbours	2.1%	5.9%	4.1%
total	83.1%	80.3%	81.3%
precision	0.82	0.77	0.79
recall	0.77	0.68	0.67
fScore	0.78	0.70	0.70
total *fScore*	0.80	0.722	0.725
SCORRACY	0.81	0.76	0.77

Table 2: A single classifier for monolithic labels vs. a combination of classifiers for separate tags. Partial matches: underspecific classifications are e.g. `s^rt` classified as `s`; overspecific — `s` classified as `s^rt`; neighbours — `s^aa` classified as `s^bk`. *Precision*, *recall* and *fScore* are means over all classifications, total *fScore* is calculated from mean *precision*/*recall*

6 Conclusion and Future Work

We have presented a new metric for the evaluation of classifiers for multidimensional dialog act tag sets—SCORRE. We have shown that such tag sets can be arranged in a hierarchical manner and that the traditional metrics *precision*, *recall* and *fScore* can be understood as distance measures in this hierarchy. SCORRE is similar to *fScore*, but does not have its asymmetric property; SCORRE is independent on the position of the labels in the hierarchy.

Future work will include further experiments, in particular how adjustments in the classifier are reflected by the SCORRE values, in order to support optimization efforts for classification results.

References

Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. 1998. Dialogue Acts in VERBMOBIL-2 Second Edition. Technical report, DFKI Saarbrücken, Universität Stuttgart, TU Berlin, Universität des Saarlandes, July.

J. Ang, Y. Liu, and E. Shriberg. 2005. Automatic dialog act segmentation and classification in multiparty meetings. In *Proc. ICASSP 2005*, Philadelphia. To appear.

A. Clark and A. Popescu-Belis. 2004. Multi-level dialogue act tags. In *Proceedings of SIGDIAL '04 (5 th SIGDIAL Workshop on Discourse and Dialog)*, Cambridge, MA.

Norbert Reithinger and Martin Klesen. 1997. Dialogue Act Classification Using Language Models. In *Proceedings of EuroSpeech-97*, pages 2235–2238, Rhodes.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. In Michael Strube and Candy Sidner, editors, *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 97–100, Cambridge, Massachusetts, USA, April 30 - May 1. Association for Computational Linguistics.