Addendum to Coding Dialogs with the DAMSL Annotation Scheme

Mark G. Core

The paper "Coding Dialogs with the DAMSL Annotation Scheme" was presented at the 1997 AAAI Fall Symposium on Communicative Actions in Humans and Machines, and can be found in the working notes for that workshop. The paper presents inter-annotator reliability results for tagging dialogs with the DAMSL annotation scheme. The results include percent pairwise agreement (PA) among annotators, percent expected agreement (PE), and kappa scores with kappa being defined as $\dfrac{PA - PE}{1 - PE}$.

This addendum presents slightly different PE results obtained by using a more accurate estimate for PE. In addition, the addendum discusses the statistical significance of the kappa scores.

Percent Expected Agreement

The tests involved 8 dialogs (for more details see the original paper); PE was calculated by computing the PE for each dialog and then calculating an average PE, weighting each PE by the number of utterances in its dialog. However, a more accurate score can be calculated by concatenating all 8 dialogs and calculating one PE. To see why this is more accurate, consider how the PE for the statement tag is calculated:

$$PE = prob(NoStatement)^2 + prob(Assert)^2 + prob(Reassert)^2 + prob(OtherStatement)^2$$

Since the probability estimates are squared, it makes more sense to calculate PE once for all the data rather than eight times (with poorer probability estimates for each of these calculations). Tables 1, 3, and 5 show the new PEs and kappas. Note, IAF is Influence on Addressee Future Action.

Significance of Kappas

Siegel and Castellan in their book, "Nonparametric Statistics for the Behavioral Sciences" show how to test the significance of kappa scores, to see whether a kappa score was a result of chance or reflects the agreement among the annotators. The equations given by Siegel and Castellan are shown below. It is assumed that kappas are normally distributed. Siegel and Castellan give a table of one tailed significance levels indexed by z values. The lowest level given is .1 so dashes in the significance level tables represent levels of higher value. The highest level given is .000005 so entries marked .000005 mean .000005 or better.

$$var(K) \approx \frac{2}{Nk(k-1)} \; \frac{PE - (2k-3)PE^2 + 2(k-2)\sum p_j^3}{(1-PE)^2}$$

$$z = \frac{K}{\sqrt{var(K)}}$$

Tables 2, 4, and 6 show the significance levels of individual dialogs (not the significance of the global scores in tables 1, 3, and 5). Because we included 3 annotators in dialog 3, it is unclear how to calculate global kappa variance since one of the parameters to variance is number of annotators. The significance levels in tables 1, 3, and 5 are based on two annotators which is close to correct since only 40 tags were contributed by the third annotator on dialog 3.

Discussion

The global kappas for Committing Speaker Future Action and Unintelligible turned out to be non-significant and are not included in the results. Unintelligible is a rare tag so it is not surprising that we do not yet have enough data for reliable statistics on it. Committing Speaker Future Action (offers and commits) is not rare as commitments occur when speakers agree to an action. However, disagreements in offers and commits make the estimate of kappa variance high giving commitment/offer data low significance. The Committing Speaker Future Action kappa reported in the paper was very low at 0.15, already indicating there were problems in the annotation manual guidelines for commit and offer.

Measure	Statement	IAF	Other For Funct
PA	0.82	0.88	0.92
PE	0.47	0.60	0.85
Kappa	0.67	0.71	0.48
Signif	0.000005	0.000005	0.000005

Table 1: Reliability for Main Forward Function Labels

Dialog	Tags	Statement	IAF	Other For Funct
d1	266	.000005	.000005	-
d2	144	.000005	.0005	-
d3	120	.000005	.000005	-
d4	82	.000005	.00005	-
d5	38	.1	-	-
d6	176	.000005	.000005	.005
d7	318	.000005	.000005	.001
d8	104	.000005	.0005	-

Table 2: Significance levels for Main Forward Function Labels

Most of the kappas (tables 1, 3, 5) are either slightly greater than the previously reported values or the same. Info-level and abandoned decreased somewhat. These changes do not alter the conclusions of the paper. A few tags such as Answer and Response-to are very reliable while others such as Agreement need improvement. The common disagreements among annotations (as discussed in the paper) are addressed in the newest annotation manual so that the next round of inter-annotator reliability tests should show that DAMSL annotated dialogs are suitable as training and testing material for dialog systems.

Acknowledgments

Thanks to Barbara Di Eugenio for suggesting these recalculations.

Measure	Understand	Agree	Ans	Resp-to
PA	0.83	0.78	0.95	0.83
PE	0.59	0.61	0.72	0.28
Kappa	0.58	0.43	0.81	0.77
Signif	0.000005	0.000005	0.000005	0.000005

Table 3: Reliability for Backward Function Labels

Dialog	Tags	Understand	Agree	Ans	Resp-to
d1	266	.000005	.00005	.00005	.000005
d2	144	.000005	.0005	.1	.000005
d3	120	.0005	-	.000005	.000005
d4	82	.005	.025	.0005	.000005
d5	38	.025	-	-	.01
d6	176	.0005	.005	.000005	.000005
d7	318	.000005	.000005	.000005	.000005
d8	104	-	.005	.0005	.000005

Table 4: Significance levels for Backward Function Labels

Measure	Info level	Abandoned
PA	0.82	0.98
PE	0.56	0.93
Kappa	0.59	0.62
Signif	0.000005	0.00005

Table 5: Reliability for Utterance Features

Dialog	Tags	Info level	Abandoned
d1	266	.00005	.005
d2	144	.005	-
d3	120	.000005	-
d4	82	.0005	-
d5	38	.025	-
d6	176	.000005	-
d7	318	.000005	.005
d8	104	.001	-

Table 6: Significance levels for Utterance Features