**Example frames for scoring**

Command: “move the two of spades to the first foundation stack”

Intended (oracle) action: the two of spades is moved to the first foundation stack

|  |  |  |
| --- | --- | --- |
| **ORACLE COMMAND**  **(ambiguous)** | **ORACLE ACTION (unambiguous)** | **PREDICTION** |
| DC= | DC= | DC= |
| FS=s | FS=s | FS=s |
| FV=2 | FV=2 | FV=1 |
| TS= | TS= | TS= |
| TV= | TV= | TV= |
| FF= | FF= | FF= |
| TF=1 | TF=1 | TF=2 |
| FC= | FC= | FC= |
| TC= | TC= | TC= |
| FH= | FH= | FH= |

Command: “move the black two to the foundation”

Intended (oracle) action: the two of clubs is moved to the first foundation stack

|  |  |  |
| --- | --- | --- |
| **ORACLE COMMAND**  **(ambiguous)** | **ORACLE ACTION (unambiguous)** | **PREDICTION** |
| DC= | DC= | DC= |
| FS=s|c | FS=c | FS=s |
| FV=2 | FV=2 | FV=1 |
| TS= | TS= | TS= |
| TV= | TV= | TV= |
| FF= | FF= | FF= |
| TF=1-4 | TF=1 | TF=1 |
| FC= | FC= | FC= |
| TC= | TC= | TC= |
| FH= | FH= | FH= |

Command: “move the black ace to the first foundation stack”

Intended (oracle) action: the ace of spades is moved to the first foundation stack

|  |  |  |
| --- | --- | --- |
| **ORACLE COMMAND**  **(ambiguous)** | **ORACLE ACTION (unambiguous)** | **PREDICTION** |
| DC= | DC= | DC= |
| FS=s|c | FS=s | FS=d |
| FV=1 | FV=1 | FV=1 |
| TS= | TS= | TS= |
| TV= | TV= | TV=2 |
| FF= | FF= | FF= |
| TF=1 | TF=1 | TF=1 |
| FC= | FC= | FC= |
| TC= | TC= | TC= |
| FH= | FH= | FH= |

Command: “move the three of spades to the foundation”

Intended (oracle) action: the three of spades is moved to the second foundation stack

|  |  |  |
| --- | --- | --- |
| **ORACLE COMMAND**  **(ambiguous)** | **ORACLE ACTION (unambiguous)** | **PREDICTION** |
| DC= | DC= | DC=1 |
| FS=s | FS=s | FS= |
| FV=3 | FV=3 | FV= |
| TS= | TS= | TS= |
| TV= | TV= | TV= |
| FF= | FF= | FF= |
| TF=1-4 | TF=2 | TF= |
| FC= | FC= | FC= |
| TC= | TC= | TC= |
| FH= | FH= | FH= |

Command: “deal card”

Intended (oracle) action: a new card is dealt

|  |  |  |
| --- | --- | --- |
| **ORACLE COMMAND**  **(ambiguous)** | **ORACLE ACTION (unambiguous)** | **PREDICTION** |
| DC=1 | DC=1 | DC= |
| FS= | FS= | FS=c |
| FV= | FV= | FV=2 |
| TS= | TS= | TS= |
| TV= | TV= | TV=2 |
| FF= | FF= | FF= |
| TF= | TF= | TF=1 |
| FC= | FC= | FC= |
| TC= | TC= | TC= |
| FH= | FH= | FH= |

Command: “deal card”

Intended (oracle) action: a new card is dealt

|  |  |  |
| --- | --- | --- |
| **ORACLE COMMAND**  **(ambiguous)** | **ORACLE ACTION (unambiguous)** | **PREDICTION** |
| DC=1 | DC=1 | DC=1 |
| FS= | FS= | FS= |
| FV= | FV= | FV= |
| TS= | TS= | TS= |
| TV= | TV= | TV= |
| FF= | FF= | FF= |
| TF= | TF= | TF= |
| FC= | FC= | FC= |
| TC= | TC= | TC= |
| FH= | FH= | FH= |

**Confusion matrices, based on the (unambiguous) oracle action frames:**

Ultimately, the objective of the system is to induce an action frame on the basis of which the correct action can be executed; that is, the action that was intended by the speaker. We therefore compute the basic scores on the basis of a comparison between the predicted frame and the oracle *action* frame: the oracle action frame corresponds to the intended action in the particular context in which it was uttered.

dealcard

|  |  |  |
| --- | --- | --- |
|  | DC=1 | DC={} |
| DC=1 | 1 | 1 |
| DC={} | 1 | 3 |

from\_suit

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | FS=s | FS=d | FS=h | FS=c | FS={} |
| FS=s | 1 | 1 |  |  | 1 |
| FS=d |  |  |  |  |  |
| FS=h |  |  |  |  |  |
| FS=c | 1 |  |  |  |  |
| FS={} |  |  |  | 1 | 1 |

from\_value

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | FV=1 | FV=2 | FV=3 | FV=4 | FV=5 | FV=6 | FV=7 | FV=8 | FV=9 | FV=10 | FV=11 | FV=12 | FV=13 | FV={} |
| FV=1 | 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FV=2 | 2 |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FV=3 |  |  |  |  |  |  |  |  |  |  |  |  |  | 1 |
| FV=4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FV=5 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FV=6 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FV=7 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FV=8 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FV=9 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FV=10 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FV=11 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FV=12 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FV=13 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| FV={} |  | 1 |  |  |  |  |  |  |  |  |  |  |  | 1 |

target\_suit

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | TS=s | TS=d | TS=h | TS=c | TS={} |
| TS=s |  |  |  |  |  |
| TS=d |  |  |  |  |  |
| TS=h |  |  |  |  |  |
| TS=c |  |  |  |  |  |
| TS={} |  |  |  |  | 6 |

target\_value

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TV=1 | TV=2 | TV=3 | TV=4 | TV=5 | TV=6 | TV=7 | TV=8 | TV=9 | TV=10 | TV=11 | TV=12 | TV=13 | TV={} |
| TV=1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| TV=2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| TV=3 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| TV=4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| TV=5 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| TV=6 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| TV=7 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| TV=8 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| TV=9 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| TV=10 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| TV=11 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| TV=12 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| TV=13 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| TV={} |  | 2 |  |  |  |  |  |  |  |  |  |  |  | 4 |

from\_foundation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | FF=1 | FF=2 | FF=3 | FF=4 | FF={} |
| FF=1 |  |  |  |  |  |
| FF=2 |  |  |  |  |  |
| FF=3 |  |  |  |  |  |
| FF=4 |  |  |  |  |  |
| FF={} |  |  |  |  | 6 |

target\_foundation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | TF=1 | TF=2 | TF=3 | TF=4 | TF={} |
| TF=1 | 2 | 1 |  |  |  |
| TF=2 |  |  |  |  | 1 |
| TF=3 |  |  |  |  |  |
| TF=4 |  |  |  |  |  |
| TF={} | 1 |  |  |  | 1 |

from\_column

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | FC=1 | FC=2 | FC=3 | FC=4 | FC=5 | FC=6 | FC=7 | FC={} |
| FC=1 |  |  |  |  |  |  |  |  |
| FC=2 |  |  |  |  |  |  |  |  |
| FC=3 |  |  |  |  |  |  |  |  |
| FC=4 |  |  |  |  |  |  |  |  |
| FC=5 |  |  |  |  |  |  |  |  |
| FC=6 |  |  |  |  |  |  |  |  |
| FC=7 |  |  |  |  |  |  |  |  |
| FC={} |  |  |  |  |  |  |  | 6 |

target\_column

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | TC=1 | TC=2 | TC=3 | TC=4 | TC=5 | TC=6 | TC=7 | TC={} |
| TC=1 |  |  |  |  |  |  |  |  |
| TC=2 |  |  |  |  |  |  |  |  |
| TC=3 |  |  |  |  |  |  |  |  |
| TC=4 |  |  |  |  |  |  |  |  |
| TC=5 |  |  |  |  |  |  |  |  |
| TC=6 |  |  |  |  |  |  |  |  |
| TC=7 |  |  |  |  |  |  |  |  |
| TC={} |  |  |  |  |  |  |  | 6 |

from\_hand

|  |  |  |
| --- | --- | --- |
|  | FH=1 | FH={} |
| FH=1 |  |  |
| FH={} |  | 6 |

|  |  |
| --- | --- |
|  | = true positive |

**Basic scores, using the (unambiguous) oracle action frames:**

**Scores using ‘overall’ formulas:**

precision = #correctly filled slots / #total filled slots in prediction = 5/16 = 0.3125

recall = #correctly filled slots / #total filled slots in reference = 5/14 = 0.3571

accuracy = (#correctly filled slots + #correct empty slots) / #total classifications = (5+40)/60 = 0.75

true negative rate = #correct empty slots / #total empty slots in reference = 40/46 = 0.8696

**Macro- and micro-averaged scores, using the *per-class* counts of true/false positives/negatives:**

***Micro-averaged precision/recall scores:***

precision = ∑i=1..MTPi / ∑i=1..M(TPi + FPi) = ∑i=1..MTPi /(∑i=1..MTPi + ∑i=1..M FPi) = 5/(5+11) = 5/16 = 0.3125

recall = ∑i=1..MTPi / ∑i=1..M(TPi + FNi) = ∑i=1..MTPi /(∑i=1..MTPi + ∑i=1..M FNi) = 5/(5+9) = 5/14 = 0.3571

(M = #classes; ∑i=1..M means sum over all classes)

* ***yields the same scores as the ‘overall’ formulas***

***Macro-averaged precision/recall scores:***

precision = ∑i=1..MPreci / M =

(1/2 + 1/3 + 0/1 + 2/3 + 0/1 + 0/2 + 1/2 + 0/1 + 0/1) / 9 = 2/9 = 0.2222

FS=s FV=1 TF=2 TF=1 FS=d TV=2 DC=1 FS=c FV=2

recall = ∑i=1..MReci / M =

(1/3 + 0/2 + 2/3 + 0/1 + 1/1 + 0/1 + 0/1 + 1/2) / 8 = 2.5/8 = 0.3125

FS=s FV=2 TF=1 FS=c FV=1 FV=3 TF=2 DC=1

**Analysis of scores, taking into account ambiguities in the commands:**

Some commands contain ambiguous expressions, which imply multiple possible values for a particular slot. For instance, in the second example command (“move the black two to the foundation”), the word “black” implies two possible values for the slot “from\_suit”: FS=s or FS=c. In addition, the expression “to the foundation” implies four possible values for the slot “target\_foundation”: TF=1, TF=2, TF=3 or TF=4.

On the basis of the command only, there is no way for the system to determine exactly which of the possible values for the slots “from\_suit” and “target\_foundation” are intended by the speaker. We could therefore argue, that on the basis of the command only, the prediction FS=s is a correct prediction, even though the intended “from\_suit” was in this case FS=c. The only problem is that the system did not induce *all* possible values implied by the command. However, the prediction of multiple values for one slot is not possible in the current implementation (see the discussion on the next page), so we could argue that the system should not be penalized for only inducing one of the possible values.

**Adapted scores**

Suppose that we adapt the scores, such that all the possible values implied by ambiguous expressions in the commands are allowed in the prediction. This would mean, that in cases where the predicted slot value does *not* correspond to the reference slot value in the oracle action frame, but *is* implied as a possible slot value by an ambiguous expression in the command, that predicted slot value would be regarded as a correct value. In the confusion matrix, this would count as a true positive for the predicted value (because the reference value in the oracle frame might as well have been the predicted value, on the basis of the command).

For the example frames above, this means that the confusion “FS=c --> FS=s” in the second frame would be changed into “FS=s --> FS=s”.

So the confusion matrix for the slot from\_suit would be changed as follows:

original

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | FS=s | FS=d | FS=h | FS=c | FS={} |
| FS=s | 1 | 1 |  |  | 1 |
| FS=d |  |  |  |  |  |
| FS=h |  |  |  |  |  |
| FS=c | 1 |  |  |  |  |
| FS={} |  |  |  | 1 | 1 |

↓

adapted

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | FS=s | FS=d | FS=h | FS=c | FS={} |
| FS=s | 2 | 1 |  |  | 1 |
| FS=d |  |  |  |  |  |
| FS=h |  |  |  |  |  |
| FS=c |  |  |  |  |  |
| FS={} |  |  |  | 1 | 1 |

For the example frames above, this would be the only adaptation in the confusion matrices.

(NB: in order to make these confusion matrices for the adapted scores, both the oracle action frames *and* the oracle command frames are needed! It cannot be done with either of the two alone.)

The resulting precision/recall scores for the example frames:

***Micro-averaged precision/recall scores:***

precision = ∑i=1..MTPi / ∑i=1..M(TPi + FPi) = ∑i=1..MTPi /(∑i=1..MTPi + ∑i=1..M FPi) = 6/(6+10) = 6/16 = 0.375

recall = ∑i=1..MTPi / ∑i=1..M(TPi + FNi) = ∑i=1..MTPi /(∑i=1..MTPi + ∑i=1..M FNi) = 6/(6+8) = 6/14 = 0.4286

***Macro-averaged precision/recall scores:***

precision = ∑i=1..MPreci / M =

(2/2 + 1/3 + 0/1 + 2/3 + 0/1 + 0/2 + 1/2 + 0/1 + 0/1) / 9 = 2.5/9 = 0.2778

FS=s FV=1 TF=2 TF=1 FS=d TV=2 DC=1 FS=c FV=2

recall = ∑i=1..MReci / M =

(2/4 + 0/2 + 2/3 + 1/1 + 0/1 + 0/1 + 1/2) / 7 = 2.66667/7 = 0.3810

FS=s FV=2 TF=1 FV=1 FV=3 TF=2 DC=1

**Implications of the two scores (basic and adapted) for the final application**

In the current system implementation, the system induces only one possible action frame, with unambiguous slot values. If that action frame does not correspond to the oracle action frame, the action taken by the system is incorrect in that particular game state, even though it might be a correct possibility according to the (potentially ambiguous) semantics of the expressed command. Therefore, our basic scoring method counts the confusion “FS=c --> FS=s“ in the second example frame as an error (even though the expression “black” in the command does include FS=s as a possible value).

In order to correctly induce the semantic frame of the command, the system would actually have to produce the ambiguous slot value FS=s|c. It would then have to determine which of the two values was appropriate in the state of the game at that particular moment, by interacting with the application. However, producing an ambiguous slot value such as FS=s|c in the output is in principle infeasible, since ambiguous classes do not occur in the training data. A feasible alternative is that the system produces an n-best list of induced frames, and then goes through that n-best list until it encounters an action frame that corresponds to a possible move in the game. In that framework, the word “black” in the command may result in an n-best list where both frames with FS=s *and* frames with FS=c have high rankings.

In the current implementation, the resulting frame is the 1-best frame. We could therefore assume that in an n-best implementation, this induced frame would be at the top of the n-best list. However, we do not know what the rest of the n-best list would look like. So if the frame induced by the current system predicts the value FS=s instead of FS=c on the basis of the word “black”, we do not know whether the correct action frame, with FS=c, would be high enough in the n-best list to be selected on the basis of the interaction with the game. Therefore, the adapted score as calculated above does not really tell us anything about the performance of a future n-best system.