	QE			TIDES		
k	`pred`	`mean`	`rand`	`pred`	`mean`	`rand`
40%	1.14 $\pm 2.9 \cdot 10^{-2}$	0.78 $\pm 6.6 \cdot 10^{-3}$	1.45	—	—	—
30%	0.94 $\pm 2.9 \cdot 10^{-2}$	0.78 $\pm 7.4 \cdot 10^{-3}$	1.44	0.95 $\pm 2.7 \cdot 10^{-2}$	0.43 $\pm 2.6 \cdot 10^{-2}$	1.37
20%	0.77 $\pm 3.4 \cdot 10^{-2}$	0.78 $\pm 1.0 \cdot 10^{-2}$	1.45	0.76 $\pm 2.6 \cdot 10^{-2}$	0.41 $\pm 2.5 \cdot 10^{-2}$	1.38
10%	0.65 $\pm 2.1 \cdot 10^{-2}$	0.79 $\pm 1.9 \cdot 10^{-2}$	1.47	0.48 $\pm 3.0 \cdot 10^{-2}$	0.41 $\pm 2.5 \cdot 10^{-2}$	1.36

Table 1: Completion performance as evaluated by the MAE for the three prediction methods and the three corpora considered.

random samples of the examples: the number of disagreement falls from 20% (Sect. 3) to less than 4%. While the `mean` method outperforms the `pred` method, this result shows that, even in case of low inter-rater agreement, there is still enough information to predict the score of one annotator knowing only the score of the others.

For the tasks considered, decisions based on a recovered matrix are therefore more similar to decisions made considering the full score matrix than decisions based on a single rating of each example.

5 Conclusion

This paper proposed a new way of collecting reliable human assessment. We showed, on two corpora, that knowing multiple scores for each example instead of a single score results in a more reliable estimation of the quality of a NLP system. We proposed to used matrix completion techniques to reduce the annotation effort required to collect these multiple ratings. Our experiments showed that while scores predicted using these techniques are pretty different from the true scores, decisions considering them are more reliable than decisions based on a single score.

Even if it can not predict scores accurately, we believe that the connection between NLP evaluation and matrix completion has many potential applications. For instance, it can be applied to identify errors made when collecting scores by comparing the predicted and actual scores.

6 Acknowledgments

This work was partly supported by ANR project Trace (ANR-09-CORD-023). The author would like to thank François Yvon and Nicolas Pécheux for their helpful questions and comments on the various drafts of this work.

% missing data	`pred`	`mean`
30%	9.24%	3.53 %
20%	6.45%	2.10 %
10%	3.66%	1.20 %

Table 3: Disagreements in a pairwise comparison of two systems of the TIDES corpus, when the systems are evaluated using the predicted scores and the true scores

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Comput. Linguist.*, 34(4):555–596, December.

Stephen R. Becker, Emmanuel J. Candès, and Michael C. Grant. 2011. Templates for convex cone problems with applications to sparse signal recovery. *Math. Prog. Comput.*, 3(3):165–218.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2012. Findings of the 2012 workshop on statistical machine translation. In *Proc. of WMT*, pages 10–51, Montréal, Canada, June. ACL.

Emmanuel Candès and Benjamin Recht. 2012. Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111–119, June.

A. Chistov and D. Grigor'ev. 1984. Complexity of quantifier elimination in the theory of algebraically closed fields. In M. Chytil and V. Koubek, editors, *Math. Found. of Comp. Science*, volume 176, pages 17–31. Springer Berlin / Heidelberg.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proc. WMT*, pages 102–121, New York City, June. ACL.

Philipp Koehn. 2012. Simulating human judgment in machine translation evaluation campaigns. In *Proc. of IWSLT*.

Charles L. Lawson and Richard J. Hanson. 1974. *Solving Least Squares Problems*. Prentice Hall.