On the Predictability of Human Assessment: when Matrix Completion Meets NLP Evaluation

Guillaume Wisniewski
Université Paris Sud
LIMSI–CNRS
Orsay, France
guillaume.wisniewski@limsi.fr

Abstract

This paper tackles the problem of collecting reliable human assessments. We show that knowing multiple scores for each example instead of a single score results in a more reliable estimation of a system quality. To reduce the cost of collecting these multiple ratings, we propose to use matrix completion techniques to predict some scores knowing only scores of other judges and some common ratings. Even if prediction performance is pretty low, decisions made using the predicted score proved to be more reliable than decision based on a single rating of each example.

1 Introduction

Human assessment is often considered as the best, if not the only, way to evaluate 'subjective' NLP tasks like MT or speech generation. However, human evaluations are doomed to be noisy and, sometimes, even contradictory as they depend on individual perception and understanding of the score scale that annotators generally use in remarkably different ways (Koehn and Monz, 2006). Moreover, annotation is known to be a long and frustrating process and annotator fatigue has been identified as another source of noise (Pighin et al., 2012).

In addition to defining and enforcing stricter guidelines, several solutions have been proposed to reduce the annotation effort and produce more reliable ratings. For instance, to limit the impact of the score scale interpretation, in the WMT evaluation campaign (Callison-Burch et al., 2012), annotators are asked to rank translation hypotheses from best to worst instead of providing absolute scores (e.g. in terms of adequacy or fluency). Generalizing this approach, several works (Pighin et al., 2012; Lopez, 2012) have defined novel annotation protocols to reduce the number of judgments that need to be collected. However, all these methods suffer from several limitations: first, they provide no interpretable information about the quality of the system (only a relative comparison between two systems is possible); second, (Koehn, 2012) has recently shown that the ranking they induce is not reliable.

In this work, we study an alternative approach to the problem of collecting reliable human assessments. Our basic assumption, motivated by the success of ensemble methods, is that having several judgments for each example, even if they are noisy, will result in a more reliable decision than having a single judgment. An evaluation campaign should therefore aim at gathering a *score matrix*, in which each example is rated by all judges instead of having each judge rate only a small subset of examples, thereby minimizing redundancy. Obviously, the former approach requires a large annotation effort and is, in practice, not feasible. That is why, to reduce the number of judgments that must be collected, we propose to investigate the possibility of using matrix completion techniques to recover the entire score matrix from a sample of its entries. The question we try to answer is whether the missing scores of one judge can be predicted knowing only scores of other judges and some shared ratings.

The contributions of this paper are twofold: i) we show how knowing the full score matrix instead of a single score for each example provides a more reliable estimation of a system quality (Section 3); ii) we present preliminary experiments

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 137–142,
Sofia, Bulgaria, August 4-9 2013. ©2013 Association for Computational Linguistics

showing that missing data techniques can be used to recover the score matrix from a sample of its entries despite the low inter-rater agreement (Section 4).

2 Matrix Completion

The recovering of a matrix from a sampling of its entries is a task of considerable interest (Candès and Recht, 2012). It can be used, for instance, in recommender systems: rows of the matrix represent users that are rating movies (columns of the matrix); the resulting matrix is mostly unknown (each user only rates a few movies) and the task consists in completing the matrix so that movies that any user is likely to like can be predicted.

Matrix completion generally relies on the *low rank hypothesis*: because of hidden factors between the observations (the columns of the matrix), the matrix has a low rank. For instance, in recommender systems it is commonly believed that only a few factors contribute to an individual's tastes. Formally, recovering a matrix $\mathbf{M}$ amounts at solving:

$$\begin{array}{ll} \text{minimize} & \text{rank}\ \mathbf{X} \\ \text{subject to} & X_{ij} = M_{ij} \quad (i,j) \in \Omega \end{array} \tag{1}$$

where $\mathbf{X}$ is the decision variable and Ω is the set of known entries. This optimization problem seeks the simplest explanation fitting the observed data.

Solving the rank minimization problem has been proved to be NP-hard (Chistov and Grigor'ev, 1984). However several convex relaxations of this program have been proposed. In this work, we will consider the relaxation of the rank by the nuclear norm[1] that can be efficiently solved by semidefinite programming (Becker et al., 2011). This relaxation enjoys many theoretical guarantees with respect to the optimality of its solution (under mild assumptions its solution is also the solution of the original problem), the conditions under which the matrix can be recovered and the number of entries that must be sampled to recover the original matrix. In our experiments we used TFOCS,[2] a free implementation of this method.

3 Corpora

For our experiments we considered two publicly available corpora in which multiple human ratings (i.e. scores on an ordinal scale) were available.

The CE Corpus The first corpus of human judgments we have considered has been collected for the WMT12 shared task on quality estimation (Callison-Burch et al., 2012).[3] The data set is made of $2,254$ English sentences and their automatic translations in Spanish predicted by a standard Moses system. Each sentence pair is accompanied by three estimates in the range 1 to 5 of its translation quality expressed in terms of post-editing effort. These human grades are in the range 1 to 5, the latter standing for a very good translation that hardly requires post-editing, while the former identifies very poor automatic translations that are not deemed to be worth the post-editing effort.

As pointed out by the task organizers, despite the special care that was taken to ensure the quality of the data, the inter-raters agreement was much lower than what is typically observed in NLP tasks (Artstein and Poesio, 2008): the weighted κ ranged from 0.39 to 0.50 depending on the pair of annotators considered[4]; the Fleiss coefficient (a generalization of κ to multi-raters) was 0.25 and the Kendall τ_b correlation coefficient[5] between 0.64 and 0.68, meaning that, on average, two raters do not agree on the relative order of two translations almost two out of five times. In fact, as often observed for the sentence level human evaluation of MT outputs, the different judges have used the score scale differently: the second judge had a clear tendency to give more 'medium' scores than the others, and the variance of her scores was low. Because theirs distributions are different, standardizing the scores has only a very limited impact on the agreement.

If, as in many manual evaluations, each example had been rated by a single judge chosen randomly, the resulting scores would have been only moderately correlated with the average of the three scores which is, intuitively, a better estimate of the 'true' quality: the 95% confidence interval of the

[1] The nuclear norm of a matrix is the sum of its singular values; the relation between rank an nuclear norm is similar to the one between ℓ_0 and ℓ_1 norms.

[2] `http://cvxr.com/tfocs/`

[3] The corpus is available from `http://www.statmt.org/wmt12/quality-estimation-task.html`

[4] The weighted κ is a generalization of the κ to ordinal data; a linear weighting schema was used.

[5] Note that, in statistics, agreement is a stronger notion than correlation, as the former compare the actual values.

τ_b between the averaged scores and the 'sampled' score is 0.754–0.755.

TIDES The second corpus considered was collected for the DARPA TIDES program: a team of human judges provided multiple assessments of adequacy and fluency for Arabic to English and Chinese to English automatic translations.[6] For space reasons, only results on the Chinese to English fluency corpus will be presented; similar results were achieved on the other corpora.

In the considered corpus, 31 sets of automatic translations, generated by three systems, have been rated by two judges on a scale of 1 to 5. The inter-rater agreement is very low: depending on the pair of judges, the weighted κ is between -0.05 and 0.2, meaning that agreement occurs less often than predicted by chance alone. More importantly, if the ratings of a pair of judges were used to decide which is the best system among two, the two judges will disagree 36% of the time. This 'agreement' score is computed as follows: if $m_{A,i}$ is the mean of the scores given to system A by the i-th annotator, we say that there is no agreement in a pairwise comparison if $m_{A,i} > m_{B,i}$ and $m_{A,j} < m_{B,j}$, i.e. if two judges rank two systems in a different order; the score is then the percentage of agreement when considering all pairs of systems and judges.

Considering the full scoring matrix instead of single scores has a large impact: if each example is rated by a single judge (chosen randomly), the resulting comparison between the two systems will be different from the decision made by averaging the two scores of the full score matrix in almost 20% of the comparisons.

4 Experimental Results

4.1 Testing the Low-Rank Hypothesis

Matrix completion relies on the hypothesis that the matrix has a low rank. We first propose to test this hypothesis on simulated data, using a method similar to the one proposed in (Mathet et al., 2012), to evaluate the impact of noise in human judgments on the score matrix rank. Artificial ratings are generated as follows: a MT system is producing n translations the quality of which, q_i, is estimated by a continuous value, that represents, for instance, a hTER score. This value is drawn from $\mathcal{N}\left(\mu, \sigma^2\right)$. Based on this 'intrinsic' quality, two ratings, a_i and b_i, are generated according to three strategies: in the first, a_i and b_i are sampled from $\mathcal{N}\left(q_i, \theta\right)$; in the second, $a_i \sim \mathcal{N}\left(q_i + \frac{\theta}{2}, \sigma'^2\right)$ and $b_i \sim \mathcal{N}\left(q_i - \frac{\theta}{2}, \sigma'^2\right)$ and in the third, $a_i \sim \mathcal{N}\left(q_i, \sigma'^2\right)$ and the b_i is drawn from a bimodal distribution $\frac{1}{2}\left(\mathcal{N}\left(q_i - \frac{\theta}{2}, \sigma'^2\right) + \mathcal{N}\left(q_i + \frac{\theta}{2}, \sigma'^2\right)\right)$ (with $\sigma'^2 < \frac{\theta}{2}$). θ describes the noise level.

Each of these strategies models a different kind of noise that has been observed in different evaluation campaigns (Koehn and Monz, 2006): the first one describes random noise in the ratings; the second a systematic difference in the annotators' interpretation of the score scale and the third, the situation in which one annotator gives medium score while the other one tend to commit more strongly to whether she considered the translation good or bad. Stacking all these judgments results in a $n \times 2$ score matrix. To test whether this matrix has a low rank or not, we assess how close it is to its approximation by a rank 1 matrix. A well-known result (Lawson and Hanson, 1974) states that the Frobenius norm of the difference of these matrices is equal to the 2nd singular value of the original matrix; the quality of the approximation can thus be estimated by ρ, defined as the 2nd eigenvalue of the matrix normalized by its norm (Leon, 1994). Intuitively, the smaller ρ, the better the approximation.

Figure 1 represents the impact of the noise level on the condition number. As a baseline, we have also represented ρ for a random matrix. All values are averaged over 100 simulations. As it could be expected, ρ is close to 0 for small noise level; but even for moderate noise level, the second eigenvalue continue to be small, suggesting that the matrix can still be approximated by a matrix of rank 1 without much loss of information. As a comparison, on average, $\rho = 0.08$ for the CE score matrix, in spite of the low inter-rater agreement.

4.2 Prediction Performance

We conducted several experiments to evaluate the possibility to use matrix completion to recover a score matrix. Experiments consist in choosing randomly $k\%$ of the entries of a matrix; these entries are considered unknown and predicted using the method introduced in Section 2 denoted `pred` in the following. In our experiments k varies from 10% to 40%. Note that, when, as in our exper-

[6] These corpora are available from LDC under the references `ldc2003t17` and `ldc2003t18`

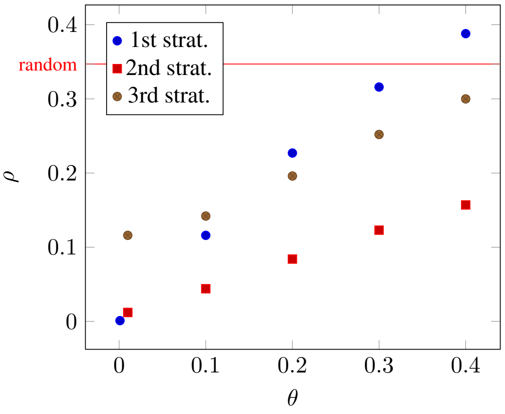

Figure 1: Evolution of the condition number ρ with the noise level θ for the different strategies (see text for details)

iments, only two judges are involved, $k = 50\%$ would mean that each example is rated by a single judge. Two simple methods for handling missing data are used as baselines: in the first one, denoted `rand`, missing scores are chosen randomly; the second one, denoted `mean`, predicts for all the missing scores of a judge the mean of her known scores.

We propose to evaluate the quality of the recovery, first by comparing the predicted score to their true value and then by evaluating the decision that will be made when considering the recovered matrix instead of the full matrix.

Prediction Performance Comparing the completed matrix to the original score matrix can be done in terms of Mean Absolute Error (MAE) defined as $\frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$ where $\hat{y}_i$ is the predicted value and y_i the corresponding 'true' value; the sum runs over all unknown values of the matrix.

Table 1 presents the results achieved by the different methods. All reported results are averaged over 10 runs (i.e.: sampling of the score matrix and prediction of the missing scores) and over all pairs of judges. All tables also report the 95% confidence interval. The MAE of the `rand` method is almost constant, whatever the number of samples is. Performance of the matrix completion technique is not so good: predicted scores are quite different than true scores. In particular, performance falls quickly when the number of missing data increases. This observation is not surprising: when 40% of the scores are missing, only a few examples have more than a single score and many have no score at all. In these conditions recovering the matrix is almost impossible. The performance of the simple `mean` technique is, comparatively, pretty good, especially when only a few entries are known. However, the `pred` method always outperform the `rand` method showing that there are dependencies between the two ratings even if statistical measures of agreement are low.

missing data	`pred`	`mean`
40%	0.78 $\pm 6.21 \times 10^{-3}$	0.72 $\pm 8.86 \times 10^{-3}$
30%	0.83 $\pm 3.19 \times 10^{-3}$	0.80 $\pm 5.42 \times 10^{-3}$
20%	0.88 $\pm 2.49 \times 10^{-3}$	0.87 $\pm 3.54 \times 10^{-3}$
10%	0.93 $\pm 1.76 \times 10^{-3}$	0.92 $\pm 1.51 \times 10^{-3}$

Table 2: Correlation between the rankings induced by the recovered matrix and the original score matrix for the CE corpus

Impact on the Decision The negative results of the previous paragraph only provide indirect measure of the recovery quality as it is not the value of the score that is important but the decision that it will support. That is why, we also evaluated matrix recovery in a more task-oriented way by comparing the decision made when considering the recovered score matrix instead of the 'true' score matrix.

For the CE corpus, a task-oriented evaluation can be done by comparing the rankings induced by the recovered matrix and by the original matrix when examples are ordered according to their averaged score. Such a ranking can be used by a MT user to set a quality threshold granting her control over translation quality (Soricut and Echihabi, 2010). Table 2 shows the correlation between the two rankings as evaluated by τ_b. The two rankings appear to be highly correlated, the matrix completion technique outperforming slightly the `mean` baseline. More importantly, even when 40% of the data are missing, the ranking induced by the true scores is better correlated to the ranking induced by the predicted scores than to the ranking induced when each example is only rated once: as reported in Section 3, the τ_b is, in this case, 0.75.

For the TIDES corpus, we computed the number of pairs of judges for which the results of a pairwise comparison between two systems is different when the systems are evaluated using the predicted scores and the true scores. Results presented in Table 3 show that considering the predicted matrix is far better than having judges rate

	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.

Stephen J: Leon. 1994. *Linear Algebra with Applications*. Macmillan,.

Adam Lopez. 2012. Putting human assessments of machine translation systems in order. In *Proc. of WMT*, pages 1–9, Montréal, Canada, June. ACL.

Yann Mathet, Antoine Widlcher, Karën Fort, Claire François, Olivier Galibert, Cyril Grouin, Juliette Kahn, Sophie Rosset, and Pierre Zweigenbaum. 2012. Manual corpus annotation: Giving meaning to the evaluation metrics. In *Proceedings of COLING 2012: Posters*, pages 809–818, Mumbai, India, December.

Daniele Pighin, Lluís Formiga, and Lluís Màrquez. 2012. A graph-based strategy to streamline translation quality assessments. In *Proc. of AMTA*.

Matthew Snover, Nitin Madnani, Bonnie Dorr, and Richard Schwartz. 2009. Fluency, adequacy, or HTER? Exploring different human judgments with a tunable MT metric. In *Proc. of WMT*, pages 259–268, Athens, Greece, March. ACL.

Radu Soricut and Abdessamad Echihabi. 2010. Trustrank: Inducing trust in automatic translations via ranking. In *Proc. of ACL*, pages 612–621, Uppsala, Sweden, July. ACL.