Hochschule Niederrhein
University of Applied Sciences

Elektrotechnik
und Informatik
Faculty of Electrical Engineering
and Computer Science

Corrigendum to: Sentiment Lexica from Paired Comparisons

Christoph Dalitz

Technischer Bericht Nr. 2017-02

IMPRESSUM

Technische Berichte des Fachbereichs Elektrotechnik und Informatik,
Hochschule Niederrhein

ISSN 2199-031X

HERAUSGEBER

Christoph Dalitz und Steffen Goebbels
Fachbereich Elektrotechnik und Informatik

ANSCHRIFT

Hochschule Niederrhein
Reinarzstr. 49
47805 Krefeld

http://www.hsnr.de/fb03/technische-berichte/

Dieses Werk wird wie folgt zitiert:

C. Dalitz: „Corrigendum to: Sentiment Lexica from Paired Comparisons.“ Technical Report No. 2017-02, Hochschule Niederrhein, Fachbereich Elektrotechnik und Informatik, 2017

Corrigendum to: Sentiment Lexica from Paired Comparisons

Christoph Dalitz
Institut für Mustererkennung
Hochschule Niederrhein
Reinarzstr. 49, 47805 Krefeld
christoph.dalitz@hsnr.de

Abstract

The sentiment scores presented by Dalitz & Bednarek in “Sentiment lexica from paired comparisons” at the ICDM Sentire workshop (2016) were based on an approximation formula by Elo that was grossly inaccurate in that particular use case. This corrigendum describes how the scores should be estimated instead and shows that these new scores are indeed a good fit to the probabilistic sentiment score model. The conclusions in the Sentire paper about the quality of the corpus based sentiment lexica SentiWS and SenticNet 3 still hold, however, because the scores obtained with the inaccurate approximation formula are similar to the correctly estimated scores when scaled with a factor, which means that the correlation is not that much affected by the error. Nevertheless, the approximate solution presented in the Sentire paper should not be used and be replaced by a numerical non-linear least squares or maximum likelihood optimization.

1 Introduction

In their presentation at the ICDM Sentire workshop [1], Dalitz & Bednarek proposed a method to assign polarity scores to words that represents the strength of the positive or negative affect associated with each word. The method uses the paired comparisons, the theory of which was originally developed in psychology [2] and later applied to chess ratings [3, 4]. The original model ignored the possibility of draws, but Dalitz & Bednarek used a generalized model that allowed draws, too [5].

Applied to word polarity, the model makes the assumption that each word w_i has a hidden rating r_i. The probability that w_i is more positive than w_j (symbolically: $w_i > w_j$) in a randomly chosen context depends on the difference between the hidden ratings:

$$P(w_i > w_j) = F(r_i - r_j - t) \tag{1a}$$

$$P(w_i \approx w_j) = F(r_i - r_j + t) - F(r_i - r_j - t) \tag{1b}$$

$$P(w_i < w_j) = F(r_j - r_i - t) \tag{1c}$$

where $(-t, t)$ is the *draw width*, and F is the cumulative distribution function of a zero-symmetric random variable. The normal distribution function is the only choice for F with a sound statistical justification (Thurstone-Mosteller model), but simpler forms for F have also been used like the logistic distribution (Bradley-Terry model) or the uniform distribution [6].

The ratings are the sentiment scores and need to be estimated from the observed comparison results. To estimate all scores, we performed a k-fold round-robin experiment from which the n unknown scores $(r_i)_{i=1}^n$ were to be estimated (case 2 in [1]). To do so, we followed the non-linear least squares estimation method by Batchelder & Bershed [3], which minimizes the squared differences between the observed scorings

$$S_i = \underbrace{W_i}_{\text{wins}} + \frac{1}{2}(\underbrace{D_i}_{\text{draws}} + \underbrace{k}_{\text{self}}) \tag{2}$$

and and their expectation values $E(S_i)$, which can be approximated by a Taylor expansion around $t = 0$ as

$$E(S_i) = k \sum_{j=1}^{n} F(r_i - r_j) + O(t^2) \tag{3}$$

The non-linear least squares estimator are the ratings $r_1, \ldots, r_n$ that minimize

$$SS(r_1, \ldots, r_n) = \sum_{i=1}^{n} \Big(S_i - k \sum_{j=1}^{n} F(r_i - r_j) \Big)^2 \tag{4}$$

In [1], we had solved Eq. (4) for its minimum analytically by making the following approximation that is due to Elo [4, paragraph 1.66]:

$$\sum_{j=1}^{n} F(r_i - r_j) \approx n \cdot F(r_i - \overline{r}) \tag{5}$$

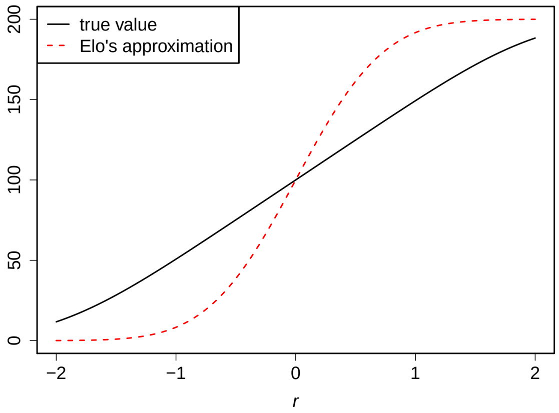

Figure 1: Comparison of Elo's approximation $n \cdot F(r - \overline{r})$ (see Eq. (5)) with the true value of $\sum_{i=1}^{n} F(r - r_i)$ as a function of r for evenly spaced r_i and a normal distribution F with $\sigma = 1/\sqrt{3}$.

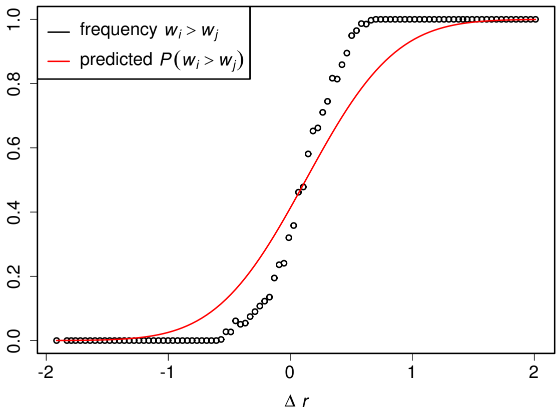

Figure 2: Comparison of the observed relative frequencies with the probabilities predicted by model (1) with a normal distribution F and scores computed with Elo's approximation.

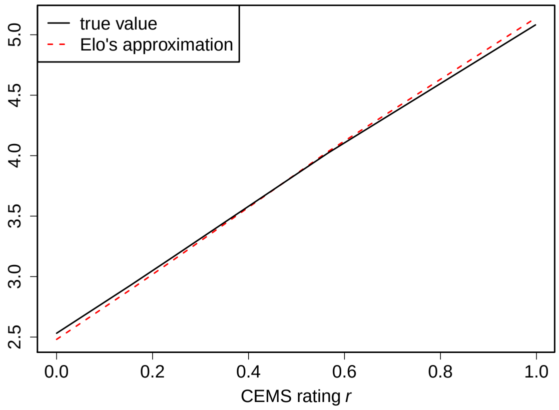

Figure 3: Comparison of Elo's approximation $n \cdot F(r - \overline{r})$ with the true value of $\sum_{i=1}^{n} F(r - r_i)$ for the CEMS data with the scores reported by Cattelan [5] for the normal distribution F and $\sigma = 1$.

where $\overline{r} = \sum_{j=1}^{n} r_i/n$ is the average rating of all words. Eq. (5) holds exactly only when F is the uniform distribution and all rating differences are within the support of the uniform distribution. In all other cases, the approximation may become inaccurate. As we will show in this corrigendum, the error is so large in this use case that the approximation must not be used and a numeric algorithm for minimizing $SS(r_1, \ldots, r_n)$ must be applied instead.

2 Inaccuracy of Elo's approximation

Let us first check directly how good Elo's approximation is in our case. We will see in the next section that the resulting scores range in our experiment is about $[-2, +2]$ for a normal distribution function F with $\sigma = 1/\sqrt{3}$. For $n = 200$ ratings equally spaced between -2 and $+2$, the values of the different sides of Eq. (5) are shown in Fig. 1.

The difference can become greater than 40 which is an error of 20% of the total possible score 200. This shows that Elo's approximation is too crude to be usable in our case. It is interesting to note that the true value is close to a linear function of r and one could get the idea to use this approximation. The slope of the line depends on the range of the scores, however, which is not known beforehand. This means that a linear approximation of $\sum_{i=1}^{n} F(r - r_i)$ cannot be used either to find the minimum of (4) analytically.

Another way to assess the quality of the approximation (5) is to compare the observed probabilities with the probabilities predicted by the model (1) with the scores obtained from the approximation through [1, Eq. (9)]. To do so, we have binned all occurring $\Delta r = r_i - r_j$ in our 2-fold round-robin experiment into 100 bins and counted the relative frequencies of $w_i > w_j$, $w_i \approx w_j$, and $w_i < w_j$ as estimators for the respective probabilities. Fig. 2 shows that the prediction is quite poor and that the scores presented at the Sentire workshop are not the best fit to the model because the score differences are estimated too small.

This raises the question why Elo's approximation worked so well in the case of the CEMS data [7], where it yielded almost the same results as the maximum-likelihood estimator. As shown in Fig. 3, the ratings are so close in this case that all differences

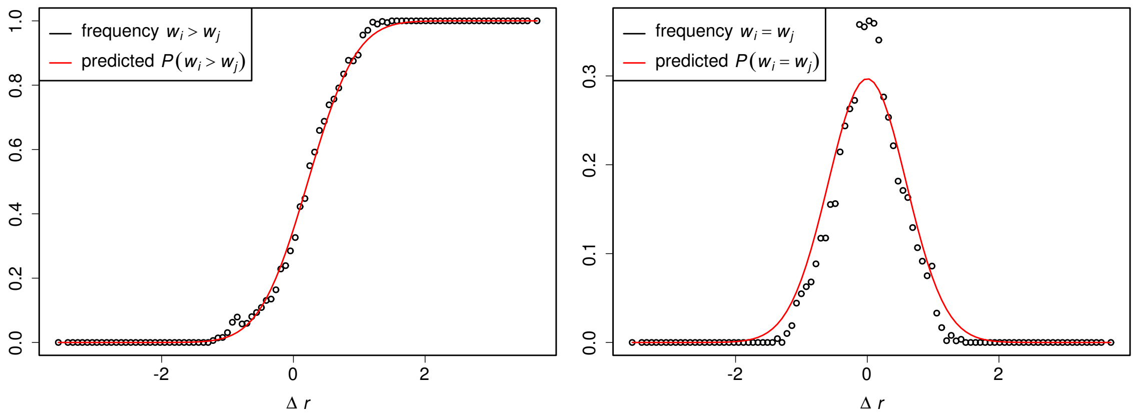

Figure 4: Comparison of the observed relative frequencies with the probabilities predicted by model (1) with a normal distribution F and scores computed by numeric minimization of (4).

fall into a range where the distribution function is almost linear. This means that Elo's approximation incidentally is quite good in this particular case. For our word sentiment score problem, however, a different estimation method must be used.

3 Correct score estimation

For obtaining correct scores, the sum of squares (4) must be minimized numerically. For non-linear least squares problems, the Levenberg-Marquardt algorithm is an efficient algorithm that is, e.g., provided by the R package *minpack.lm* [8]. As can be seen in Fig. 4, the ratings estimated with this method lead to a model in good agreement with the observed comparison results.

The range of the least squares fitted scores is about $[-2, 2]$ when σ is set to $1/\sqrt{3}$, while the range of the scores obtained with Elo's approximation was about $[-1, 1]$. The draw width t is greater, too (0.220 versus 0.128). Nevertheless, the Pearson correlation between both scores is 0.9973, and the Spearman correlation even 1.0000. This means that the scores from Elo's approximation are linearly transformed by a factor around 0.5, which theoretically could be corrected *after* score estimation by reducing the scale parameter σ in F. This means that the probability $P(\text{“unpraktisch”} > \text{“rüde”})$ reported in [1, p. 928] was too small (0.58) and is actually greater (0.64).

An alternative approach to estimate $r_1, \ldots, r_n$ would be to maximize the log-likelihood function

$$l(r_1, \ldots, r_n, t) = \sum_{\substack{\textit{comparisons}\\ \textit{with } w_i > w_j}} \log F(r_i - r_j - t) \qquad (6)$$

$$+ \sum_{\substack{\textit{comparisons}\\ \textit{with } w_i \approx w_j}} \log \Big(F(r_i - r_j + t) - F(r_i - r_j - t) \Big)$$

The resulting model fit is similar to Fig. 4, but with an even slightly wider range of score values: $[-1.955, 2.132]$ versus $[-1.832, 1.904]$. The runtime for maximum-likelihood estimation is considerably greater[1], however, and numeric optimization of (6) fails in the case of the uniform distribution, because the objective function is not differentiable and many values of the ratings lead to zero probabilities. The best fit with non-linear least squares even has $l(r_1, \ldots, r_n, t) = -\infty$. For other than the uniform distribution, the maximum-likelihood estimation is a good alternative, however, especially as it does not make the assumption of a small draw width t. In our situation it is $t \approx 0.2$, and the Taylor expansion around $t = 0$ is justified, but this might not hold in more general use cases of the paired comparison model.

4 Correlation with other lexica

In the presentation for the Sentire workshop, we had used the scores to evaluate the relative quality of the corpus based sentiment lexica SentiWS [9] and SenticNet 3 [10] by means of their Pearson correlation with the paired comparison scores. Based on these correlations, we concluded that SenticNet is in better agreement with our ground truth data. As can be seen from Table 1, our conclusion still holds with the

[1] Numeric minimization of (4) with the R function *nls.lm* took 12s on an Intel i7-4770, while it took 8min for the maximization of (6) with *optim*.

	choice for F			
	normal	*logistic*	*uniform*	
direct	0.977	0.978	0.973	*LSQ*
SentiWS	0.714	0.715	0.713	
SenticNet	0.759	0.762	0.751	
direct	0.968	0.961	0.979	*Elo*
SentiWS	0.709	0.707	0.710	
SenticNet	0.741	0.732	0.763	

Table 1: Pearson correlation r_p of the polarity scores with scores from direct assignment and corpus-based lexica. "LSQ" are the results with scores correctly estimated with non-linear least squares. For comparison, the correlations with the erroneously estimated scores from [1] are given ("Elo").

correctly estimated scores, although they have a range about twice as wide. As the correlation between the erroneous (Elo) and the correct (LSQ) scores is high, the difference in their range has less effect on their correlation with other sentiment lexica than one should have expected from the inaccuracy of Elo's approximation in this case.

There is one notable difference, however: for the correct scores, the uniform distribution no longer shows the highest correlation with the other lexica. On the contrary: it is lower, albeit only slightly. Moreover, the plot for the uniform distribution corresponding to Fig. 4 shows a slightly poorer agreement between model prediction and observed judgments. In contrary to the suggestion in [1], there is thus no reason to prefer the uniform distribution.

5 Conclusion

The approximation formula for estimating the word sentiment scores in [1] must not be used. The scores must instead be computed either by non-linear least squares minimization of Eq. (4), or by maximizing the log-likelihood function (6). This also affects the computation of scores for new words, where the estimation step in lines 20 and 24 of Algorithm 1 [1, p. 927] must be replaced with a maximum likelihood or non-linear least squares estimate.

A more general lesson can be learned from this example: always verify the approximations made in a model after fitting the model to the observed data! I am sorry that we did not do this before our Sentire presentation and that this corrigendum was necessary.

References

[1] C. Dalitz and K. E. Bednarek, "Sentiment lexica from paired comparisons," in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pp. 924–930, 2016.

[2] L. L. Thurstone, "A law of comparative judgment.," *Psychological Review*, vol. 34, no. 4, pp. 368–389, 1927.

[3] W. H. Batchelder and N. J. Bershad, "The statistical analysis of a Thurstonian model for rating chess players," *Journal of Mathematical Psychology*, vol. 19, no. 1, pp. 39–60, 1979.

[4] A. E. Elo, *The Rating of Chess Players, Past and Present.* New York: Arco, 1978.

[5] M. Cattelan, "Models for paired comparison data: A review with emphasis on dependent data," *Statistical Science*, vol. 27, no. 3, pp. 412–433, 2012.

[6] G. E. Noether, "Remarks about a paired comparison model," *Psychometrika*, vol. 25, no. 4, pp. 357–367, 1960.

[7] R. Dittrich, R. Hatzinger, and W. Katzenbeisser, "Modelling the effect of subject-specific covariates in paired comparison studies with an application to university rankings," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 47, no. 4, pp. 511–525, 1998.

[8] T. V. Elzhov, K. M. Mullen, A.-N. Spiess, and B. Bolker, *minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK*, 2016. R package version 1.2-1.

[9] R. Remus, U. Quasthoff, and G. Heyer, "SentiWS - a publicly available German-language resource for sentiment analysis," in *Conference on Language Resources and Evaluation (LREC)*, pp. 1168–1171, 2010.

[10] E. Cambria, D. Olsher, and D. Rajagopal, "SenticNet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis," in *AAAI conference on artificial intelligence*, pp. 1515–1521, 2014.