s-AWARE: Using Crowd Judgements in Supervised Measure-Based Methods for IR Evaluation*

Marco Ferrante[1], Nicola Ferro[2], and Luca Piazzon[2]

[1] Department of Mathematics "Tullio Levi-Civita", University of Padua, Italy
ferrante@math.unipd.it
[2] Department of Information Engineering, University of Padua, Italy
{ferro,piazzonl}@dei.unipd.it

Abstract. Crowdsourcing methodologies have recently emerged as a cheap and fast alternative to the traditional document assessment process for ground truth creation. Early approaches make use of voting and/or classification methodologies to combine crowd judgements into a merged pool, used as reference in the evaluation process.
A measure-based approach has instead been used in *Assessor-driven Weighted Averages for Retrieval Evaluation (AWARE)* [3], focusing in optimizing the final evaluation measure without merging judgements at pool level.
s-AWARE extends AWARE with a set of supervised methods. We rely on several TREC collections to evaluate s-AWARE and we show that it outperforms state-of-the-art methods. Moreover, our results show that when moving to the real case scenario where a crowd-assessor only judges a portion of the dataset, s-AWARE is quite an effective approach.

1 Introduction

Document assessment for ground-truth creation is one of the most demanding tasks in preparing an experimental collection in both terms of time and costs, and it has traditionally been performed by relying on expert assessors [8].
Crowdsourcing methodologies [2] have been recently exploited for a faster and cheaper collection of multiple, even less qualified, document assessments. These judgements are used together in the evaluation process with the objective of achieving a proficient evaluation, comparable to the traditional one. The most common way to use crowd-judgements is to create a merged pool to be used as the gold standard for evaluation. Since errors in the merged pool can unfairly affect evaluation measures, in our work we moved the merging process at measure level, as firstly proposed in *Assessor-driven Weighted Averages for Retrieval Evaluation (AWARE)*[3]. Performance measures are firstly computed

* Extended abstract of [4].

based on each crowd-assessor judgements and then merged weighting by an estimate of each assessor accuracy, computed making use of unsupervised estimators. s-AWARE extends AWARE and uses supervised estimators based on the closeness between each assessor and the gold standard in a small set of training topics. We evaluated our s-AWARE against the state-of-the-art supervised and unsupervised methods by using several TREC datasets, achieving promising results.

This extended abstract will describe s-AWARE methodology and performance, presenting some related works (Section 2), the description of the approaches (Section 3), the performed experiments (Section 4) and possible future extensions (Section 5).

2 Related Works

One of the first developed crowd-assessor merging approach is *Majority Vote (MV)* [11], that assigns to each document the judgement proposed by the majority of crowd-assessors; Weighted versions of MV have been proposed do boost proficient assessors, e.g. [12,11].
Expectation Maximization (EM) algorithms optimize the probability of relevance of each document in an unsupervised [6] or semi-supervised way [10] and then assign to each document the most probable judgement . Another EM alternative [5] uses a variant of the same algorithm to optimize assessor reliability to be used to weight crowd judgements.
One weakness of the above described pool merging strategies is the possibility to propagate mislabelling errors to evaluation measures. Different measures could even be differently affected by the same pool error. *Assessor-driven Weighted Averages for Retrieval Evaluation (AWARE)* tries to overcome this problem by performing evaluation on the judgements given by every crowd-assessor and combining the obtained measures weighting each assessor with his accuracy, estimated in an unsupervised way favouring assessors behaving differently from some fake random assessors:

$$aware\_\mu(r_t) = \sum_{k=1}^{m} \mu\left(\hat{r}_t^k\right) \frac{a_k(t)}{\sum_{h=1}^{m} a_h(t)}$$

where m is the number of crowd-assessors to merge, $\mu\left(\hat{r}_t^k\right)$ is the value of the performance measure computed on run r for topic t according to the k-th crowd-assessor, and a_k is the accuracy of the k-th crowd-assessor. AWARE computes accuracies as a function of the distance from random assessors: the more a crowd-assessor is far from a set of random assessors, the better it is.

3 s-AWARE Methodology

To describe s-AWARE accuracy estimation, we consider the matrix M_k containing the measures computed for a set S of systems and a set T of topics based on the judgements issued by the k-th crowd assessor, and we define M^* as the gold standard measures matrix. The idea behind s-AWARE is to assign an higher accuracy to assessors that behaved similarly to the gold standard on a set of training topics. We consider the two best performing approaches used in AWARE to quantify the "closeness" C_k to the gold standard:

– *Measure closeness*: we consider the *Root Mean Square Error (RMSE)* between the crowd-measure and the gold standard one

$$C_k = RMSE\left(\overline{M}_k(\cdot, S) - \overline{M}^*(\cdot, S)\right) = \sqrt{\sum_{s=1}^{|S|} \frac{\left(\overline{M}_k(\cdot, s) - \overline{M}^*(\cdot, s)\right)^2}{\mid S \mid}}$$

where $\overline{M}(\cdot, s)$ indicates the average measure by topic

– *Ranking of Systems closeness*: we use the Kendall's τ correlation between the ranking of systems based on the crowd-measures and the gold standard one

$$C_k = \tau\left(\overline{M}_k(\cdot, S), \overline{M}^*(\cdot, S)\right) = \frac{A - D}{\mid S \mid (\mid S \mid -1)/2}$$

where A is the number of system pairs ranked in the same order in $\overline{M}_k(\cdot, S)$ and $\overline{M}^*(\cdot, S)$, and D is the number of discordant pairs.

C_ks are then normalized in the [0,1] range, obtaining normalized C_k equal to 1 with gold standard behaviour. Squared and cubed C_k are also considered to sharpen the distinction between good and bad assessors.

4 Experiments

4.1 Setup

We compared s-AWARE approaches against MV, EM with MV seeding[6], AWARE with uniform accuracy scores (`uniform`), unsupervised AWARE (u-AWARE) `unsup_rmse_tpc` and `unsup_tau_tpc` approaches (using respectively RMSE and Kendall's τ GAP computation), Georgescu Zhu EM method (hard labels, PN discrimination, no boost version) (`emGZ`) [5] and semi-supervised EM (using 30% of the documents as training set)(`emsemi`) [10].
In our evaluation, for each approach, we evaluated the systems with *Average Precision (AP)*, and we evaluated each approach performance by computing the *AP Correlation (APC)* [15] between the ranking induced by AP values and the gold standard ranking.
We used two different collections, using the NIST judgments as gold standard:

– *TREC 2012 Crowdsourcing track* [9]: 31 complete pools of judgements on 10 topics common to TREC 08 Adhoc track (`T08`) [14] and TREC 13 Robust track(`T13`) [13]. We used the 129 runs from T08 and the 110 runs from T13.
– *TREC 2017 Common Core track* (`T26`) [1]: real crowdsourced judgements gathered by Inel Et al. [7] from 406 crowd assessors on short documents ($\leq$ 1000 words) within the NYTimes corpus. Judgements refer to 50 topics, having exactly 7 judgements for each (topic, document) pair. We used the 75 runs from T26.

We tested s-AWARE using only the 30% of the topics as training set. We considered k-tuples from 2 to 7 crowd-assessors. We validated the results by repeating both topic and assessor sampling 100 times for each k-tuples size.

We performed experiments under two possible scenarios, considering *Whole Assessors* and *Partitioned Assessors*. In the *Whole Assessors* case (most favorable to supervised approaches but quite unrealistic) each crowd-assessor completely judges all the topics. *Whole Assessors* data is available only for the `T08` and `T13` tracks. In the *Partitioned Assessors* case (real case scenario, more challenging for supervised approaches), each crowd-assessor judges just a portion of the documents for a portion of the topics. Therefore, to get the complete pools assigned to each *Partitioned Assessor* we group judgements coming from different crowd-assessors. This is the case of `T26` track, but we also simulated this configuration on the `T08` and `T13` tracks, by assembling the judgments coming from more participants into each topic.

4.2 Main Results

Table 1 reports the AP Correlation results in the tested configurations on the test portion of the dataset (70% of the documents from 70% of the topics, the common subset of documents unseen by both s-AWARE and `emsemi`). The best performing approach in the *Whole Assessors* case is our `sup_tau_cubed`, constantly achieving better performance with respect to all the other approaches. More in general, as expected, s-AWARE approaches generally outperform the baselines and the corresponding unsupervised u-AWARE approach, that anyway significantly outperform the baselines.
We notice a very poor performance of `emGZ` and only a little improvement of `emsemi` with respect to `emmv`. This is probably due to the very limited amount of training data, more effectively exploited by s-AWARE.
In *Partitioned Assessors* case we face up a different situation, where s-AWARE advantage is limited with respect to u-AWARE approaches. On `T08` and `T13`, `unsup_rmse_tpc` u-AWARE method performs generally better than s-AWARE, but s-AWARE still outperform the other u-AWARE approaches and the baselines. This narrower gap supports the idea that the *Partitioned Assessors* case is less favorable to supervised approaches, since the training phase reflects less what happens in the test phase; In general, we can observe that s-AWARE still performs remarkably better than `emsemi`.
Looking to `T26`, s-AWARE approaches always outperform all the other approaches, with `sup_tau_cubed` achieving the best performance for all k-tuples. This is very promising since, while Partitioned assessors for `T08` and `T13` are simulated, `T26` is the only dataset obtained by real crowd assessors, showing a good performance in a real case scenario. In fact, we hypothesize that bad performance on `T08` and `T13` can be due to the little more fragmentation of the simulated partitioned assessors, i.e. smaller pieces from more crowd-assessors, with respect to the the `T26` ones.
In all our results, Kendall's τ performs better than RMSE as s-AWARE "closeness" accuracy computation, and cubed and squared s-AWARE approaches achieve,

in general, better performance than the basic closeness approach, since they emphasize more sharply the difference between good and bad assessors. Moreover, results highlight that s-AWARE approaches can obtain good results even with small k-tuple size.

		sup_rmse	sup_tau	sup_rmse_squared	sup_tau_squared	sup_rmse_cubed	sup_tau_cubed	unsup_rmse_tpc	unsup_tau_tpc	uniform	mv	emmv	emGZ	emsemi
T08-whole	k02	0.6048	0.6184	0.6086	0.6278	0.6120	**0.6326**	0.6075	0.6031	0.6008	0.5326	0.5183	0.5455	0.5470
	k03	0.6317	0.6499	0.6366	0.6659	0.6414	**0.6766**	0.6324	0.6298	0.6265	0.6099	0.6025	0.5413	0.6097
	k04	0.6492	0.6707	0.6546	0.6905	0.6598	**0.7045**	0.6422	0.6501	0.6436	0.6147	0.6154	0.5562	0.6329
	k05	0.6689	0.6958	0.6751	0.7221	0.6812	**0.7409**	0.6808	0.6732	0.6625	0.6569	0.6512	0.5445	0.6535
	k06	0.6555	0.6833	0.6620	0.7120	0.6685	**0.7340**	0.6622	0.6651	0.6492	0.6163	0.5918	0.5095	0.5963
	k07	0.6719	0.6998	0.6782	0.7274	0.6845	**0.7482**	0.6709	0.6834	0.6657	0.6696	0.6396	0.5028	0.6443
T13-whole	k02	0.6111	0.6192	0.6139	0.6238	0.6162	**0.6254**	0.6005	0.6078	0.6079	0.5410	0.4974	0.5012	0.5186
	k03	0.6526	0.6616	0.6562	0.6692	0.6594	**0.6733**	0.6254	0.6548	0.6486	0.6088	0.5926	0.4770	0.6085
	k04	0.6687	0.6825	0.6728	0.6941	0.6765	**0.7008**	0.6250	0.6823	0.6641	0.6214	0.6119	0.4910	0.6241
	k05	0.7061	0.7237	0.7106	0.7387	0.7148	**0.7478**	0.6797	0.7209	0.7011	0.6613	0.6491	0.4478	0.6497
	k06	0.6872	0.7068	0.6923	0.7253	0.6971	**0.7379**	0.6502	0.7151	0.6818	0.6197	0.5913	0.4289	0.5919
	k07	0.7045	0.7232	0.7092	0.7402	0.7135	**0.7515**	0.6552	0.7330	0.6996	0.6708	0.6452	0.4062	0.6476
T08-part	k02	0.5314	0.5390	0.5332	0.5456	0.5350	0.5500	**0.5508**	0.5317	0.5294	0.4919	0.4944	0.5024	0.4913
	k03	0.5466	0.5587	0.5497	0.5700	0.5526	0.5783	**0.5831**	0.5457	0.5436	0.5171	0.5292	0.5050	0.5321
	k04	0.5549	0.5690	0.5584	0.5830	0.5621	0.5935	**0.6037**	0.5553	0.5512	0.5153	0.4967	0.4992	0.5191
	k05	0.5564	0.5725	0.5604	0.5891	0.5645	0.6019	**0.6168**	0.5599	0.5523	0.5368	0.4804	0.4914	0.5118
	k06	0.5683	0.5863	0.5729	0.6064	0.5775	0.6226	**0.6552**	0.5692	0.5638	0.5287	0.4785	0.4782	0.4962
	k07	0.5672	0.5900	0.5737	0.6150	0.5797	0.6333	**0.6872**	0.5696	0.5615	0.5373	0.4774	0.4639	0.4776
T13-part	k02	0.5842	0.5959	0.5862	0.6038	0.5879	**0.6078**	0.5998	0.5767	0.5820	0.5406	0.5052	0.4945	0.4847
	k03	0.6155	0.6299	0.6181	0.6406	0.6206	**0.6474**	0.6412	0.6015	0.6126	0.5728	0.5854	0.4611	0.5742
	k04	0.6372	0.6528	0.6402	0.6647	0.6430	**0.6722**	0.6706	0.6270	0.6340	0.5848	0.5757	0.4157	0.5838
	k05	0.6481	0.6641	0.6515	0.6773	0.6549	0.6862	**0.6929**	0.6508	0.6444	0.6079	0.5619	0.3521	0.6009
	k06	0.6616	0.6776	0.6653	0.6914	0.6691	0.7015	**0.7211**	0.6663	0.6579	0.6165	0.5573	0.3044	0.5840
	k07	0.6560	0.6728	0.6603	0.6884	0.6642	0.7006	**0.7306**	0.6412	0.6512	0.6209	0.5332	0.1963	0.5568
T26-part	k02	0.3817	0.4008	0.3796	0.4084	0.3774	**0.4124**	0.3531	0.3928	0.3837	0.3731	0.3362	0.3506	0.3625
	k03	0.3863	0.4067	0.3839	0.4151	0.3815	**0.4191**	0.3522	0.4028	0.3886	0.3783	0.3512	0.3753	0.3680
	k04	0.3824	0.4072	0.3795	0.4179	0.3767	**0.4236**	0.3421	0.4029	0.3853	0.3791	0.3525	0.3688	0.3625
	k05	0.3832	0.4102	0.3796	0.4228	0.3761	**0.4295**	0.3396	0.4077	0.3866	0.3785	0.3602	0.3648	0.3729
	k06	0.3926	0.4232	0.3896	0.4366	0.3870	**0.4441**	0.3568	0.4207	0.3961	0.3781	0.3584	0.3466	0.3737
	k07	0.4534	0.4787	0.4521	0.4918	0.4507	**0.4980**	0.4171	0.4841	0.4561	0.4400	0.4302	0.3715	0.4239

Table 1: AP Correlation results. Baseline approaches are in blue, u-AWARE in green, s-AWARE in orange. Darker color indicate better performance. Best performing approaches for each k-tuple size are in bold.

5 Conclusions and Future Work

We presented s-AWARE [4] a methodology for merging crowd-assessors, that extends AWARE approach to supervised techniques. We tested s-AWARE against a set of unsupervised and supervised baselines, highlighting the effectiveness of s-AWARE in the very challenging real scenario situation where only 30% of the documents were used for training. s-AWARE outperform all the others in the *Whole Assessors* case and is still quite robust in the *Partitioned Assessors* case. In the future, we plan to extend AWARE framework to better deal with partial assessments, assigning an accuracy score to each real crowd assessor, avoiding the need to group judgements as done in Partitioned assessor case.

References

1. Allan, J., Harman, D.K., Kanoulas, E., Li, D., Van Gysel, C., Voorhees, E.M.: TREC 2017 Common Core Track Overview. In: Voorhees, E.M., Ellis, A. (eds.) The Twenty-Sixth Text REtrieval Conference Proceedings (TREC 2017). National Institute of Standards and Technology (NIST), Special Publication 500-324, Washington, USA (2018)
2. Alonso, O.: The Practice of Crowdsourcing. Morgan & Claypool Publishers, USA (May 2019)
3. Ferrante, M., Ferro, N., Maistro, M.: AWARE: Exploiting Evaluation Measures to Combine Multiple Assessors. ACM Transactions on Information Systems **36**(2), 1–38 (aug 2017). https://doi.org/10.1145/3110217, `https://doi.org/10.1145%2F3110217`
4. Ferrante, M., Ferro, N., Piazzon, L.: s-aware: Supervised measure-based methods for crowd-assessors combination. In: Arampatzis, A., Kanoulas, E., Tsikrika, T., Vrochidis, S., Joho, H., Lioma, C., Eickhoff, C., Névéol, A., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 16–27. Springer International Publishing, Cham (2020)
5. Georgescu, M., Zhu, X.: Aggregation of crowdsourced labels based on worker history. In: Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14). WIMS '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2611040.2611074, `https://doi.org/10.1145/2611040.2611074`
6. Hosseini, M., Cox, I.J., Milić-Frayling, N., Kazai, G., Vinay, V.: On aggregating labels from multiple crowd workers to infer relevance of documents. In: Proceedings of the 34th European Conference on Advances in Information Retrieval. pp. 182–194. ECIR'12 (2012)
7. Inel, O., Haralabopoulos, G., Li, D., Van Gysel, C., Szlávik, Z., Simperl, E., Kanoulas, E., Aroyo, L.: Studying Topical Relevance with Evidence-based Crowdsourcing. In: Cuzzocrea, A., Allan, J., Paton, N.W., Srivastava, D., Agrawal, R., Broder, A., Zaki, M.J., Candan, S., Labrinidis, A., Schuster, A., Wang, H. (eds.) Proc. 27th International Conference on Information and Knowledge Management (CIKM 2018). pp. 1253–1262. ACM Press, New York, USA (2018)
8. Sanderson, M.: Test Collection Based Evaluation of Information Retrieval Systems. Foundations and Trends in Information Retrieval (FnTIR) **4**(4), 247–375 (2010)
9. Smucker, M.D., Kazai, G., Lease, M.: Overview of the TREC 2012 Crowdsourcing Track. In: Voorhees, E.M., Buckland, L.P. (eds.) The Twenty-First Text REtrieval Conference Proceedings (TREC 2012). National Institute of Standards and Technology (NIST), Special Publication 500-298, Washington, USA (2013)
10. Tang, W., Lease, M.: Semi-supervised consensus labeling for crowdsourcing. In: Proceedings of the SIGIR 2011 workshop on crowdsourcing for information retrieval (CIR). pp. 36–41. Association for Computing Machinery, New York, NY, USA (2011)
11. Tao, D., Cheng, J., Yu, Z., Yue, K., Wang, L.: Domain-weighted majority voting for crowdsourcing. IEEE Transactions on Neural Networks and Learning Systems **30**(1), 163–174 (jan 2019). https://doi.org/10.1109/tnnls.2018.2836969, `https://doi.org/10.1109%2Ftnnls.2018.2836969`
12. Tian, T., Zhu, J., Qiaoben, Y.: Max-margin majority voting for learning from crowds. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(10), 2480–2494 (oct 2019). https://doi.org/10.1109/tpami.2018.2860987, `https://doi.org/10.1109%2Ftpami.2018.2860987`

13. Voorhees, E.M.: Overview of the TREC 2004 Robust Track. In: Voorhees, E.M., Buckland, L.P. (eds.) The Thirteenth Text REtrieval Conference Proceedings (TREC 2004). National Institute of Standards and Technology (NIST), Special Publication 500-261, Washington, USA (2004)
14. Voorhees, E.M., Harman, D.K.: Overview of the Eigth Text REtrieval Conference (TREC-8). In: Voorhees, E.M., Harman, D.K. (eds.) The Eighth Text REtrieval Conference (TREC-8). pp. 1–24. National Institute of Standards and Technology (NIST), Special Publication 500-246, Washington, USA (1999)
15. Yilmaz, E., Aslam, J.A., Robertson, S.E.: A New Rank Correlation Coefficient for Information Retrieval. In: Chua, T.S., Leong, M.K., Oard, D.W., Sebastiani, F. (eds.) Proc. 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008). pp. 587–594. ACM Press, New York, USA (2008)