Unsupervised System Combination for Set-based Retrieval with Expectation Maximization*

Han-Chin Shing,[2] Joe Barrow,[2] Petra Galuščáková,[1]
Douglas W. Oard,[1,3] and Philip Resnik[1,4]

[1]UMIACS/[2]Comp.Sci./[3]iSchool/[4]Linguistics, University of Maryland, College Park
{shing,jdbarrow}@cs.umd.edu, petra@umiacs.umd.edu, {oard,resnik}@umd.edu

Abstract. System combination has been shown to improve overall performance on many rank-based retrieval tasks, often by combining results from multiple systems into a single ranked list. In contrast, set-based retrieval tasks call for a technique to combine results in ways that require decisions on whether each document is in or out of the result set. This paper presents a set-generating unsupervised system combination framework that draws inspiration from evaluation techniques in sparse data settings. It argues for the existence of a duality between evaluation and system combination, and then capitalizes on this duality to perform unsupervised system combination. To do this, the framework relies on the consensus of the systems to estimate latent "goodness" for each system. An implementation of this framework using data programming is compared to other unsupervised system combination approaches to demonstrate its effectiveness on CLEF and MATERIAL collections.

Keywords: Unsupervised System Combination · Expectation-Maximization.

1 Introduction

System combination, or data fusion, has been shown to improve performance over individual systems across a variety of information retrieval (IR) tasks [1, 8, 12, 18]. Most of the literature focuses on rank-based retrieval, where the goal is to generate a merged and improved ranked list. Returning the entire ranked list, however, might not always be optimal. Set-based retrieval studies the situation in which returning a subset of the entire rank list can be beneficial, such as when the downstream application involves heavy computation (e.g, question answering, summarization, or machine translation).

System combination for set-based retrieval has not been as well studied as for rank-based retrieval. It is also important to note that our interest in a set as the final retrieval result does not, however, mean that it must be sets that we take as the input to the combination process: a set selection strategy (such as finding a cutoff or a threshold) can be applied after merging ranked retrieval systems. This leads to questions about whether it is best to do set selection before the combination, after the combination, or even both. Furthermore, set selection often requires training data or expert knowledge and can vary greatly depending on the query of interest.

* This work has been supported in part by IARPA/AFRL contract FA8650-17-C-9117.

In this paper, we present a fully unsupervised set-generating system combination technique. It yields competitive results with widely used rank-based system combination methods that require further tuning data for set selection. Our approach draws inspiration from a duality between evaluation and system combination: evaluation allows us to estimate how good an underlying system is, and system combination can benefit from estimating how good each system is. By using expectation-maximization (E-M), we can estimate latent relevance labels with which to evaluate the systems, and then combine the systems based on how good each system is estimated to be (if we treat those labels as correct).

2 Related Work

A duality between system combination and evaluation has been suggested in the evaluation literature. Soboroff et al. [19] show that if you randomly assign relevance judgements and use them to evaluate the systems in TREC, you can still get a ranking of systems that correlates fairly well with the official ranking. Nuray and Can [16] further show that using system combination results as pseudo-relevance judgements can increase this correlation. However, they do not report how well the pseudo-relevance judgements correlate with the truth.

E-M techniques for ranked-retrieval system combination have been explored by Klementiev et al. [10]. In their work on unsupervised rank aggregation, they used the extended Mallows model to estimate the quality of each ranker's output by comparing it to a "consensus ranking". Although Klementiev et al. developed a method for unsupervised rank aggregation, their method does not address set-based retrieval, which would require further tuning a threshold to cut off the merged rank list. Other unsupervised rank aggregation methods, such as Borda counts [2], reciprocal rank fusion [3], or CombMNZ [18], all require threshold tuning when sets are the goal.

Data Programming [17] introduces an alternative E-M framework for generating a large pseudo-gold collection by combining many simple rules constructed by experts. Its E-M framework estimates the goodness of each rule by considering credit assignment and evaluation jointly, though the focus is on constructing a labelled collection to train a representation learning model. In this paper, we adapt the model presented in the Data Programming work to a set-based retrieval setting. We show that the model naturally fits the problem of set-based retrieval and allows us to combine diverse rankers.

As our interest in set-based retrieval stems from reducing the load on computationally intensive downstream applications such as QA, missing a relevant document has a high cost. Two recall-oriented metrics are thus used: F_3, and *Actual Query Weighted Value* (AQWV) [13]. F_3 is the well known F-measure with a heavier weight placed on recall. AQWV is a measure that combines the recall, $P_{\text{q,recall}}$, and the probability of a false alarm, $P_{\text{q,false\_alarm}}$:

$$\text{AQWV} = Avg_{q \in Q_{rel}} P_{\text{q,recall}} - \zeta \cdot Avg_{q \in Q} P_{\text{q,false\_alarm}} \tag{1}$$

$Q_{rel} \subset Q$ is the set of queries that has any relevant document. Since we usually have many more irrelevant documents than relevant documents, $P_{\text{q,false\_alarm}}$ is

usually quite small compared to $P_{\text{q,recall}}$. Thus, $\zeta \propto \frac{N_{\text{total}}}{N_{\text{relevant}}}$ is used to control the balance between $P_{\text{q,recall}}$ and $P_{\text{q,false\_alarm}}$.[1]

Keyword Specific Thresholding (KST) [9] is a set selection and score normalization method originally developed for the Spoken Term Detection task on a measure with a similar structure to AQWV [6]. KST is designed to find an optimal threshold for each query when performing set-based retrieval. By assuming that the score is the probability that the retrieval is correct, KST calculates a threshold for each query by balancing the risk of miss and false alarm using Bayes decision theory. For each query q, a threshold ρ_q is calculated as:

$$\rho_q = \frac{\zeta N_{q,relevant}}{\|D\| + (\zeta - 1)N_{q,relevant}} \tag{2}$$

However, as $N_{q,relevant}$ is unknown at test time, it is estimated by summing over all the scores of documents retrieved by a query, $N_{q,relevant} = \delta \sum_{d=1}^{D} s_{q,d}^{\gamma}$, where δ and γ are tunable parameters. We use $\delta = 1.5$ and $\gamma = 1$ as suggested by Wang and Metze [20].

3 System Combination Model

The proposed system combination model consists of two components: a method of set selection for the underlying rankers based on KST, and an E-M algorithm for combination. We are given a document collection, D, a set of queries, Q, and a set of rankers to combine, M. We then let the raw score assigned by ranker $m \in M$ to the pair (query $q \in Q$, document $d \in D$) be $s_{m,q,d}$. We use s_m as a shorthand for $s_{m,q,d}$ if the query and document can be inferred from context.

Set Selection The threshold ρ_q calculated from KST (see Equation 2) is used to partition the documents into three sets:

$$s'_{m,q,d} = \begin{cases} 1 & \text{if } s_{m,q,d} > \rho_q \\ -1 & \text{if } s_{m,q,d} \text{ not available (i.e., } d \text{ is not retrieved)} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $s'_{m,q,d}$ is the normalized score. As a ranker might not fully assign scores to all documents, -1 is assigned to any document that is missing a score, whereas $s'_{m,q,d} = 0$ represents that ranker m retrieved the document, but is not confident about the results.

Expectation-Maximization. Suppose a "goodness" measure for each ranker is provided a priori, then we would know something about how to combine results from the rankers. Conversely, if we use the system combination results as pseudo-gold labels to evaluate the rankers we wish to combine, we can estimate the "goodness" of each ranker, which we can further use to combine the rankers again. This formulation can be captured by E-M [5]. In this paper, we use Data Programming [17] to estimate two "goodness" measures: accuracy, α_m,. and

[1] For the MATERIAL Somali and Swahili collections, $\zeta = 40$. For the CLEF French collection, $\zeta = 240$ to account for the fact that $\frac{N_{\text{total}}}{N_{\text{relevant}}}$ for CLEF French is 6 times more than that of MATERIAL Somali or Swahili. See Table 1.

	MATERIAL				CLEF	
	Swahili		Somali		French	
	Dev	Eval	Dev	Eval	Dev	Eval
Query #	300	1,000	300	1,000	194	141
Doc #	666	14,745	695	15,377	87,191	90,261
Relevance #	434	20,198	540	17,247	3,413	3,185

Table 1: Counts of queries, documents, and positive relevance judgments.

coverage, β_m. Coverage is the probability that a ranker will assign a **confident** label (i.e. $s'_m \in \{1, -1\}$). Accuracy is the probability that a ranker is **correct** given that it is **confident**. We can derive the following probability distribution:

$$\mu_{\boldsymbol{\alpha},\boldsymbol{\beta}}(S', y) = \tfrac{1}{2}\prod_{m=1}^{M}\left(\alpha_m\beta_m\mathbb{I}_{[s'_m=y]} + (1-\alpha_m)\beta_m\mathbb{I}_{[s'_m\neq y]} + (1-\beta_m)\mathbb{I}_{[s'_m=0]}\right) \quad (4)$$

where $S' = \left\{s'_1, s'_2, \cdots, s'_{\|M\|}\right\}$ represents a collection of the rankers' normalized scores across a single document and query pair, y is the latent relevance of said document and query pair, and $\boldsymbol{\alpha} = \{\alpha_1, \cdots, \alpha_{\|M\|}\}, \boldsymbol{\beta} = \{\beta_1, \cdots, \beta_{\|M\|}\}$ are the collections of parameters from each ranker.

We wish to find $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ such that they maximize:

$$\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}} = arg\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \sum_{d=i}^{D} log\left(\sum_{y=\{-1,1\}} \mu_{\alpha,\beta}\left(S', y\right)\right) \quad (5)$$

Note that the latent relevance, y, is marginalized in the above equation, so both the E and M steps can be combined into Equation 5. Also note that the above maximization sums over all the log probabilities of the documents over a single query, and thus our E-M combination is a per-query combination technique.

Finally, to obtain the probability of retrieval for each query/document pair, we can calculate:

$$p\left(y = 1|\boldsymbol{\alpha}, \boldsymbol{\beta}, S'\right) = \frac{\mu_{\boldsymbol{\alpha},\boldsymbol{\beta}}\left(S', y = 1\right)}{\sum_{y'=\{-1,1\}} \mu_{\boldsymbol{\alpha},\boldsymbol{\beta}}\left(S', y'\right)}. \quad (6)$$

If $p\left(y = 1|\boldsymbol{\alpha}, \boldsymbol{\beta}, S'\right) > 0.5$, a (query, doc) pair is considered relevant.

4 Experimental Setup

We report results on the IARPA MATERIAL Swahili and Somali collections and the CLEF French collection [15], which are Cross-Language Information Retrieval (CLIR) collections. All three collections contain relevance judgments for English queries and documents in different languages, see Table 1. Each collection consists of two disjoint sets of sub-collections: a development collection and an evaluation collection.[2]

We choose four underlying CLIR rankers based on diversity: three Document Translation methods done with machine translation (two Neural MT [7, 14], one Statistical MT [11]) and one Query Translation [4] method.

[2] For MATERIAL Swahili and Somali, development and evaluation collections are provided by IARPA. For CLEF French, we selected query sets 2000-2003 with document sets ATS 94, Le Monde 94 as the development collection, and query sets 2004-2006 with document sets ATS 95, Le Monde 95 as the evaluation collection.

Baselines. We compare our results against three rank-based system combination methods: (1) RR sums the reciprocal rank (RR) of the documents across all rankers to re-rank the documents [3], (2) BORDA sums the $N - Rank$ of the document, where N is the maximum rank across rankers [2], and (3) COMBMNZ [18]. For CombNMZ, a standard score normalization technique is used before applying combination [12]: $CombMNZ_{d,q} = t \cdot \sum_{m=1}^{M} \frac{s_m - \min s_m}{\max s_m - \min s_m}$,where t is the number of times $s_{m,d,q}$ has a value across the rankers.

Scaled and Oracle: Note that although the baselines can generate a merged ranked list without supervision, they still require a set selection process that involves tuning on the collection, as our goal is set-based retrieval. We report two set selection approaches: (1) SCALED: tune rank cutoffs for each of the combination methods on the development collection, and project the cutoffs to the evaluation collection by multiplying by $\frac{\|D_{\text{EVAL}}\|}{\|D_{\text{DEV}}\|}$. (2) ORACLE: to remove the confounding error of selecting a cutoff and effectiveness of system combination, we also report results at the *oracle* rank cutoff tuned on the *evaluation* collections. In all cases, the same learned rank cutoff is applied for every query.

Expectation-Maximization. EM - as described in Section 3. Each ranker to be combined is first normalized by KST using Equation 3 with the default parameters. Then the E-M combination is applied to combine the rankers.

5 Results and Discussion

Table 2 shows the results of the three evaluation collections. Overall, EM performs well, achieving the best scores on Somali and French by either set-based measure. On Swahili, EM again achieves the highest F_3, and its AQWV is exceeded only by our ORACLE baseline. Notably, our EM method on Somali outperforms our ORACLE baselines with rank cutoffs unfairly (against ourselves) optimized on the *evaluation* collection. This is possible in part due to a score cutoff technique like KST calculating different thresholds for each query, whereas a tied rank cutoff is used in the ORACLE baselines. We also note that there are substantial differences between the performance of the SCALED and ORACLE baselines on Swahili and Somali; that likely results from scaling an integer cutoff on the small development set by a large factor (≈ 22.13 in each case, see Table 1). More generally, this suggests that tuning on small development collections may be useful, but sometimes far from optimal. When the development and evaluation collections have similar size, as in French, the difference is not as evident. On the other hand, our EM method, without using training data, demonstrates robustness and competitiveness across different collections.

6 Conclusion

We have presented an unsupervised set-generating system combination technique. Drawing inspiration from the evaluation literature, we reason that a duality exists between evaluation and system combination. We show that E-M combination, by jointly solving the problems of credit assignment and threshold selection, can be both effective and robust in a low resource setting where relevance judgments that can be used for development are limited or nonexistent.

		Somali		Swahili		French	
		F_3	AQWV	F_3	AQWV	F_3	AQWV
	EM	**18.47**	**19.00**	**28.09**	29.54	**39.30**	**47.66**
Scaled	CombMNZ	14.62	15.93	22.92	27.55	35.22	44.29
	RR	14.75	14.59	24.31	27.39	36.48	46.73
	Borda	13.82	13.60	20.59	26.05	34.84	43.05
Oracle	CombMNZ	17.98	16.73	26.28	30.04	35.58	45.09
	RR	17.85	16.18	26.44	**30.89**	36.92	**47.62**
	Borda	17.22	13.94	24.44	26.68	35.03	43.51

Table 2: Results on Evaluation collections.

Finally, using two set-based retrieval measures, we compare with both scaled and oracle versions of the baselines and show that E-M combination achieves competitive results.

References

1. Belkin, N., et al.: Combining the evidence of multiple query representations for information retrieval. IP&M **31**(3), 431–448 (1995)
2. de Borda, J.C.: Mémoire sur les élections au scrutin (1781)
3. Cormack, G., et al.: Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: SIGIR. vol. 9, pp. 758–759 (2009)
4. Darwish, K., Oard, D.: Probabilistic structured query methods. In: SIGIR (2003)
5. Dempster, A., et al.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society **39**(1), 1–22 (1977)
6. Fiscus, J.G., Ajot, J., Garofolo, J.S., Doddington, G.: Results of the 2006 spoken term detection evaluation. In: SIGIR. vol. 7, pp. 51–57 (2007)
7. Haddow, B., et al.: The University of Edinburgh's submissions to the WMT18 news translation task. In: WMT. pp. 403–413. Belgium, Brussels (October 2018)
8. Harman, D.: Overview of the third text retrieval conference. In: TREC (1995)
9. Karakos, D., et al.: Score normalization and system combination for improved keyword spotting. In: ASRU. pp. 210–215 (2013)
10. Klementiev, A., Roth, D., Small, K.: Unsupervised rank aggregation with distance-based models. In: ICML. pp. 472–479 (2008)
11. Koehn, P., et al.: Moses: Open source toolkit for statistical machine translation. In: ACL. pp. 177–180 (2007)
12. Lee, J., et al.: Analyses of multiple evidence combination. In: SIGIR (1997)
13. NIST: The Official Original Derivation of AQWV (2017), https://www.nist.gov/sites/default/files/documents/2017/10/26/aqwv_derivation.pdf
14. Niu, X., et al.: Bi-directional differentiable input reconstruction for low-resource neural machine translation. CoRR **abs/1811.01116** (2018)
15. Nunzio, G.M.D., et al.: CLEF 2006: Ad hoc track overview. In: CLEF (2006)
16. Nuray, R., Can, F.: Automatic ranking of information retrieval systems using data fusion. IP&M **42**(3), 595–614 (2006)
17. Ratner, A., et al.: Data programming: Creating large training sets, quickly. In: NIPS. pp. 3567–3575 (2016)
18. Shaw, J., Fox, E.: Combination of multiple searches. In: TREC (1994)
19. Soboroff, I., et al.: Ranking retrieval systems without relevance judgments. In: SIGIR. pp. 66–73 (2001)
20. Wang, Y., Metze, F.: An in-depth comparison of keyword specific thresholding and sum-to-one score normalization. In: ISCA (2014)