Unsupervised Ensemble of Ranking Models for News Comments Using Pseudo Answers

Soichiro Fujita [1], Hayato Kobayashi [2], Manabu Okumura [1]

1. Tokyo Institute of Technology

2. Yahoo Japan Corporation / RIKEN AIP

Background

Task: Ranking comments on online news services

Goal: Display high quality comments

Problem: `` high quality'' has complex factors

Problem: Ranking news comments is difficult

- We have various situations of judging whether a comment is good
  - Indicating rare user experiences
  - Providing new ideas
  - Causing discussion
- A single ranking model often fail to capture these aspects

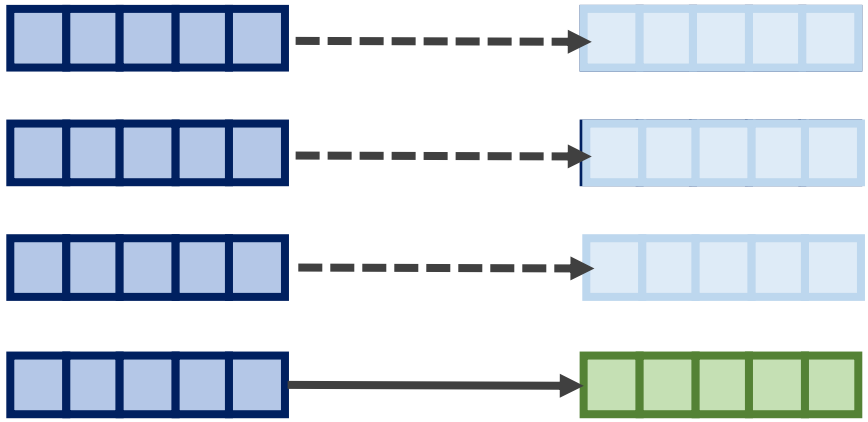
How to deal with this problem? → Ensemble techniques

- Ensemble techniques combine multiple model outputs
- Different models could capture different aspects

Two basic ensemble techniques

Selecting

Model outputs Selected output

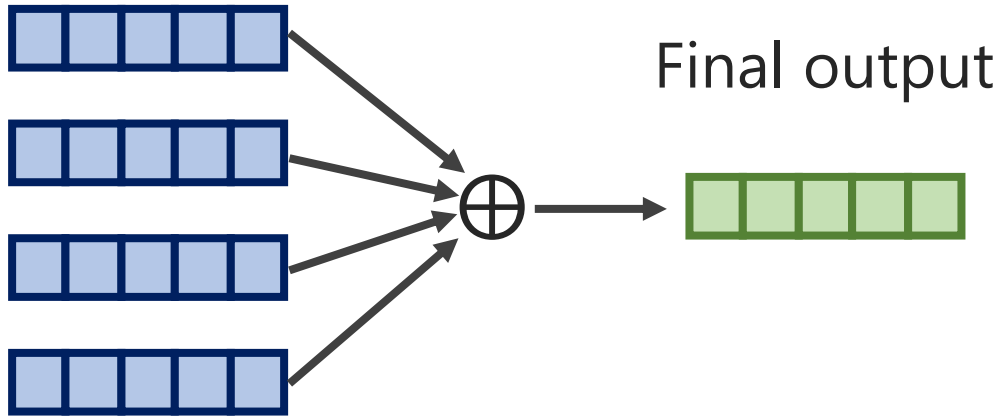


✓ Denoising lower accuracy models

✗ Depend on a single model output

Averaging

Model outputs



✓ Make up for other models' mistakes

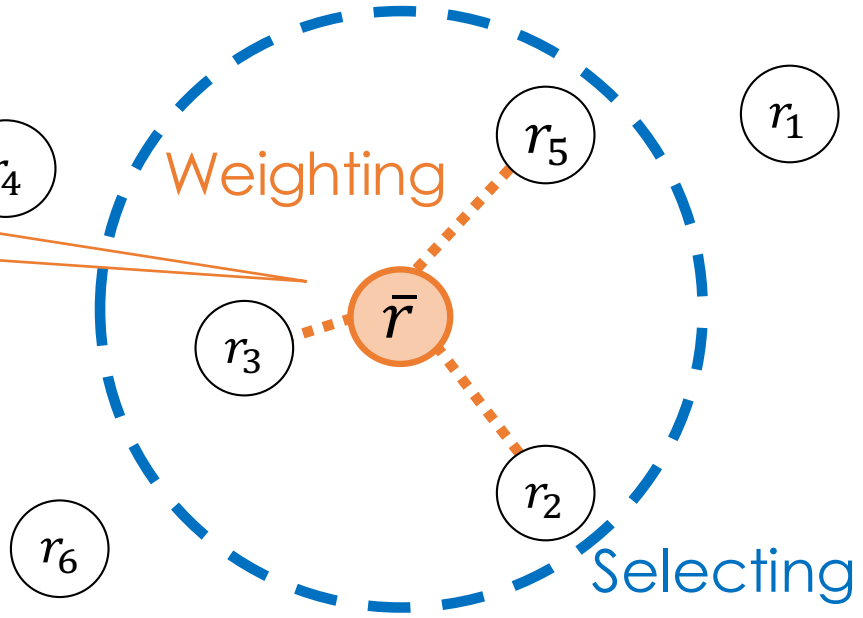
✗ Lower accuracy models could be noise

Proposed method: HPA ensemble

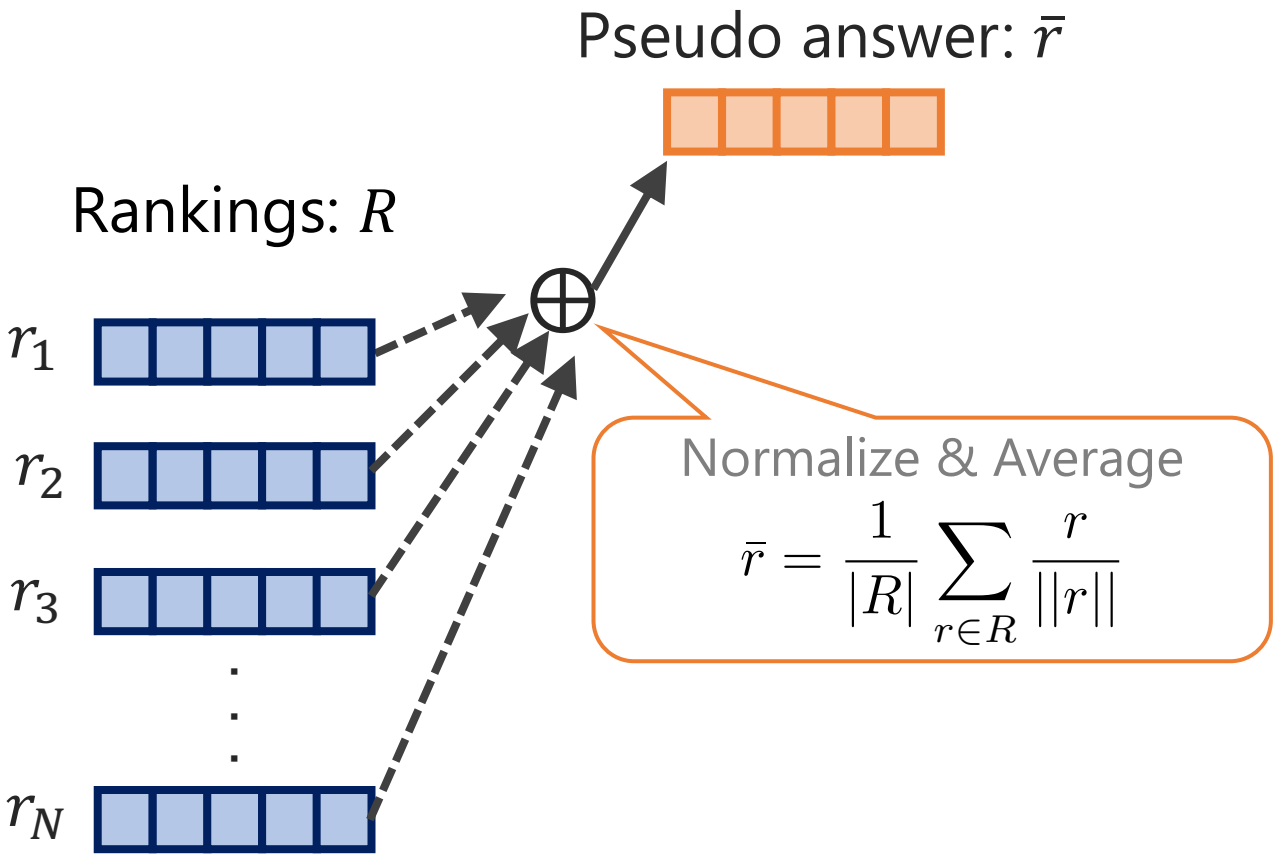
- **H**ybrid method using the **P**seudo **A**nswer
  - Hybrid of an output selection and a typical averaging method
  - Dynamic denoising of outputs via a pseudo answer $\bar{r}$

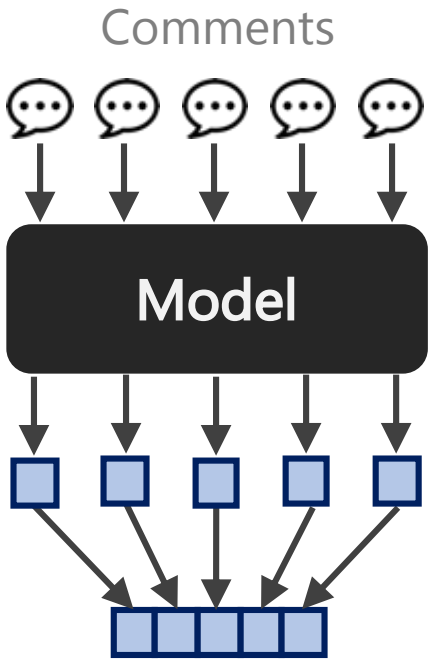
Combine only accurate outputs

Remove lower accuracy outputs



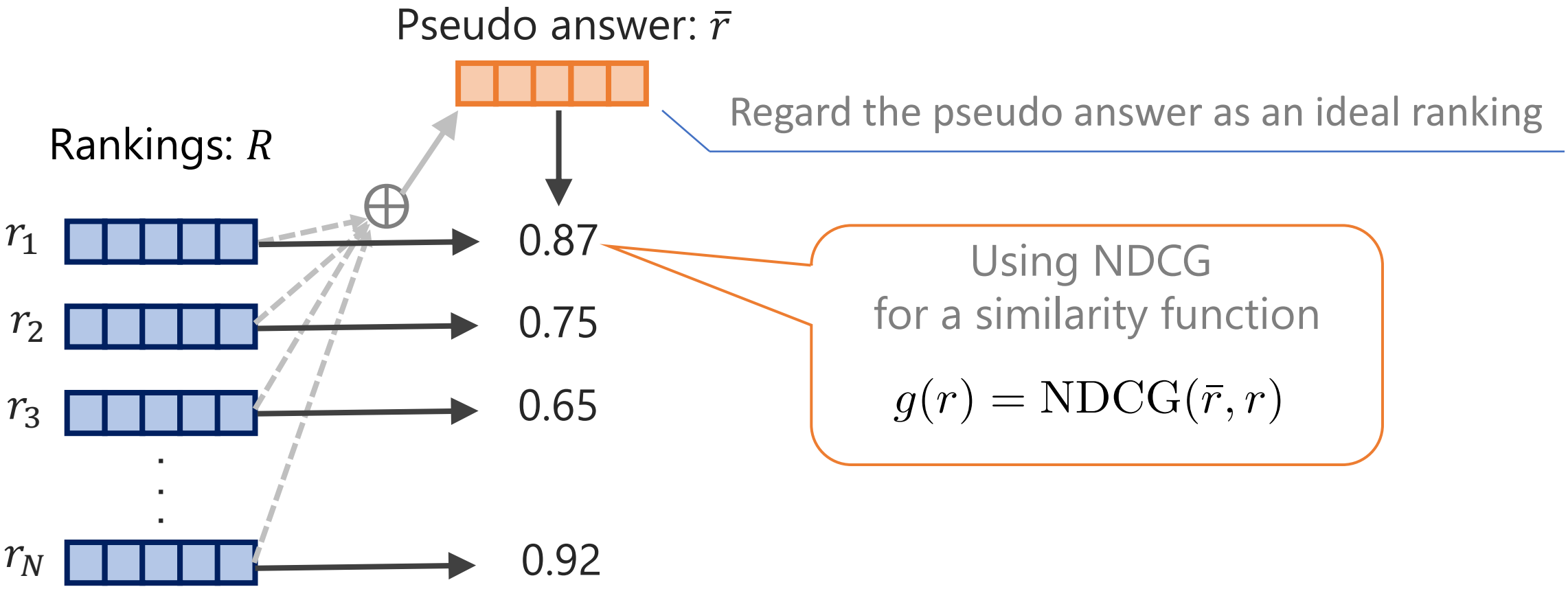
Step1: Calculate a pseudo answer



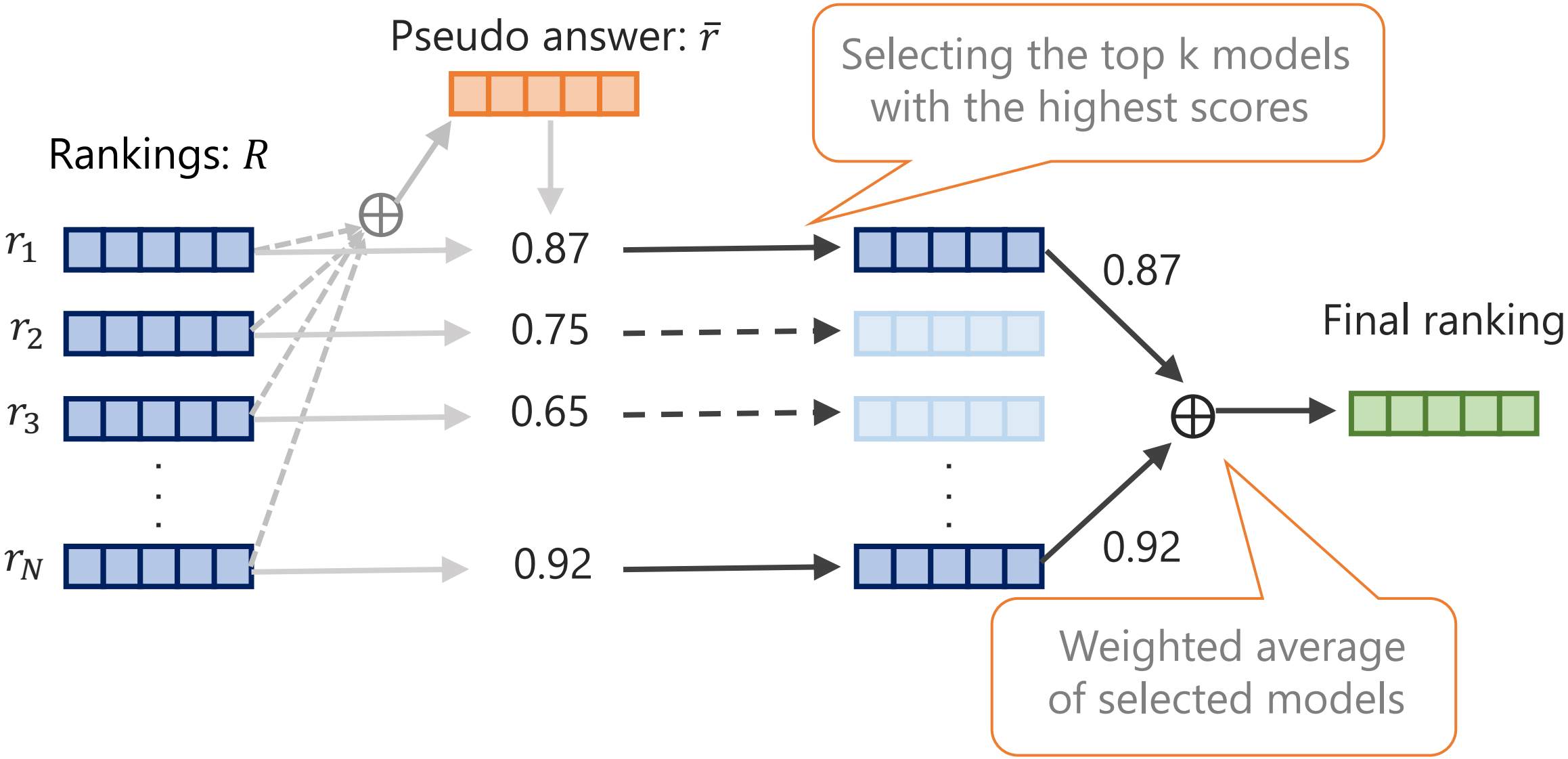


Each element represents a ranking score of a comment

Step2: Calculate similarity scores of each predicted ranking



Step3: Calculate the final ranking from similarity scores



Experimental settings

Dataset: **YJ Constructive Comment Ranking Dataset**

Train 1,300 articles, Validation 113 articles, Test 200 articles

(each article associated with more than 100 comments)

Dataset Link

Models: **LSTM-based RankNet**

Prepared 100 different models by random initialization

Metrics: **NDCG@k** and **Precision@k** ($k \in \{1, 5, 10\}$)

Evaluation results

	Methods	NDCG @1	NDCG @5	NDCG @10	Prec. @1	Prec. @5	Prec. @10
Best single model	RankNet	76.35	77.97	79.52	15.0	33.20	42.99
Unsupervised baseline	NormAvg	**79.83**	**80.77**	**82.16**	**17.08**	**37.18**	46.48
Supervised baseline	SupWeight	78.64	80.33	81.94	16.28	35.47	**46.58**
Ours	HPA	79.87	81.43	82.33	17.08	37.39	47.34
Ours w/o weighting	SPA	79.68	80.96	82.19	17.08	35.87	46.68
Ours w/o selecting	WPA	79.87	81.39	82.17	17.08	37.88	46.63

This is a part of the results.
Please see Table 1 in our paper if you want to find other baselines.

Evaluation results

	Methods	NDCG @1	NDCG @5	NDCG @10	Prec. @1	Prec. @5	Prec. @10
Best single model	RankNet	76.35	77.97	79.52	15.0	33.20	42.99
Unsupervised baseline	NormAvg	**79.83**	**80.77**	**82.16**	**17.08**	**37.18**	46.48
Supervised baseline	SupWeight	78.64	80.33	81.94	16.28	35.47	**46.58**
Ours	HPA	79.87	81.43	82.33	17.08	37.39	47.34
Ours w/o weighting	SPA	79.68	[illegible]	82.19	17.08	35.87	46.68
Ours w/o selecting							

Our method achieved the best performance

Evaluation results

	Methods	NDCG @1	NDCG @5	NDCG @10	Prec. @1	Prec. @5	Prec. @10
Best single model	RankNet	76.35	77.97	79.52	15.0	33.20	42.99
Unsupervised baseline	NormAvg	**79.83**	**80.77**	**82.16**	**17.08**	**37.18**	46.48
Supervised baseline	SupWeight	78.64	80.33	81.94	16.28	35.47	**46.58**
Ours	HPA	79.87	81.43	82.33	17.08	37.39	47.34
Ours w/o weighting	SPA	79.68	80.96	82.19	17.08	35.87	46.68
Ours w/o selecting	WPA	79.87	81.39	82.17	17.08	37.88	46.63

Hybrid of weighting and selecting is effective

Conclusion

Proposed Method:

- A hybrid unsupervised method using pseudo answers

Result:

- Our method achieved the best performance
- Denoising predicted rankings using the pseudo answer is effective

Future work:

- Combine various types of network structures
- Investigate effectiveness of our methods on other ranking datasets