YJRS at the NTCIR-14 OpenLiveQ-2 Task

June 13, 2019

Tomohiro Manabe, Sumio Fujita, Akiomi Nishida
Yahoo Japan Corporation

Abstract

6 simple approaches:

- CA (training method)+basic (features)
- CA+BM25F
- CA+translated
- CA+all
- ListNet+basic
- LightGBM+basic

Evaluation findings:

- 4 approaches passed the first round.
- ListNet+basic was quite promising.

Our Approaches

Our Approaches

They are combinations of 3 ranking models (or training methods) and 4 feature sets.

Run IDs	Ranking Model	Training Method	Feature Set
91, 95	Linear Combination	Coordinate Ascent	Basic
92			Basic + BM25F-like
93			Basic + translated
136			All
100, 113	Neural Ranking	ListNet	Basic
144, 148	Ensemble Trees	LightGBM	Basic

CA + Basic

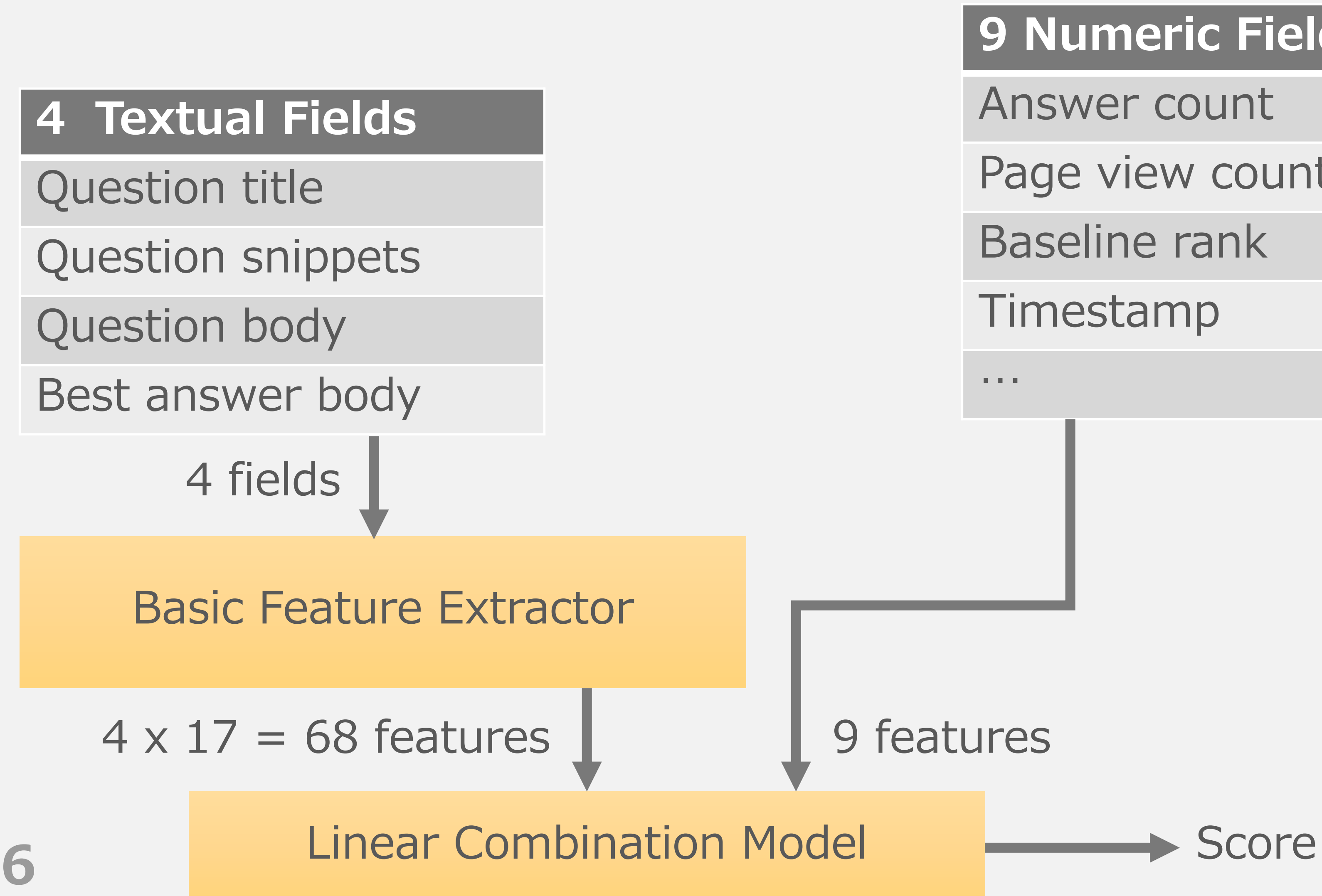
It optimizes a linear combination of 77 features with Coordinate Ascent (CA).

- 17 LETOR features of 4 textual fields
  - TF, TFIDF, BM25, field length, ⋯
- 9 numeric fields
  - Answer count, page view count, baseline rank, timestamp, ⋯

Refs: Qin *et al*. LETOR: A Benchmark Collection for Research on Learning to Rank for Information Retrieval, *Info. Retrieval*, 2010.

README of the official tool
https://github.com/mpkato/openliveq/blob/master/README.md

CA + Basic



9 Numeric Fields
Answer count
Page view count
Baseline rank
Timestamp
…
4 Textual Fields
Question title
Question snippets
Question body
Best answer body
4 fields
Basic Feature Extractor
4 x 17 = 68 features
9 features
Linear Combination Model
Score

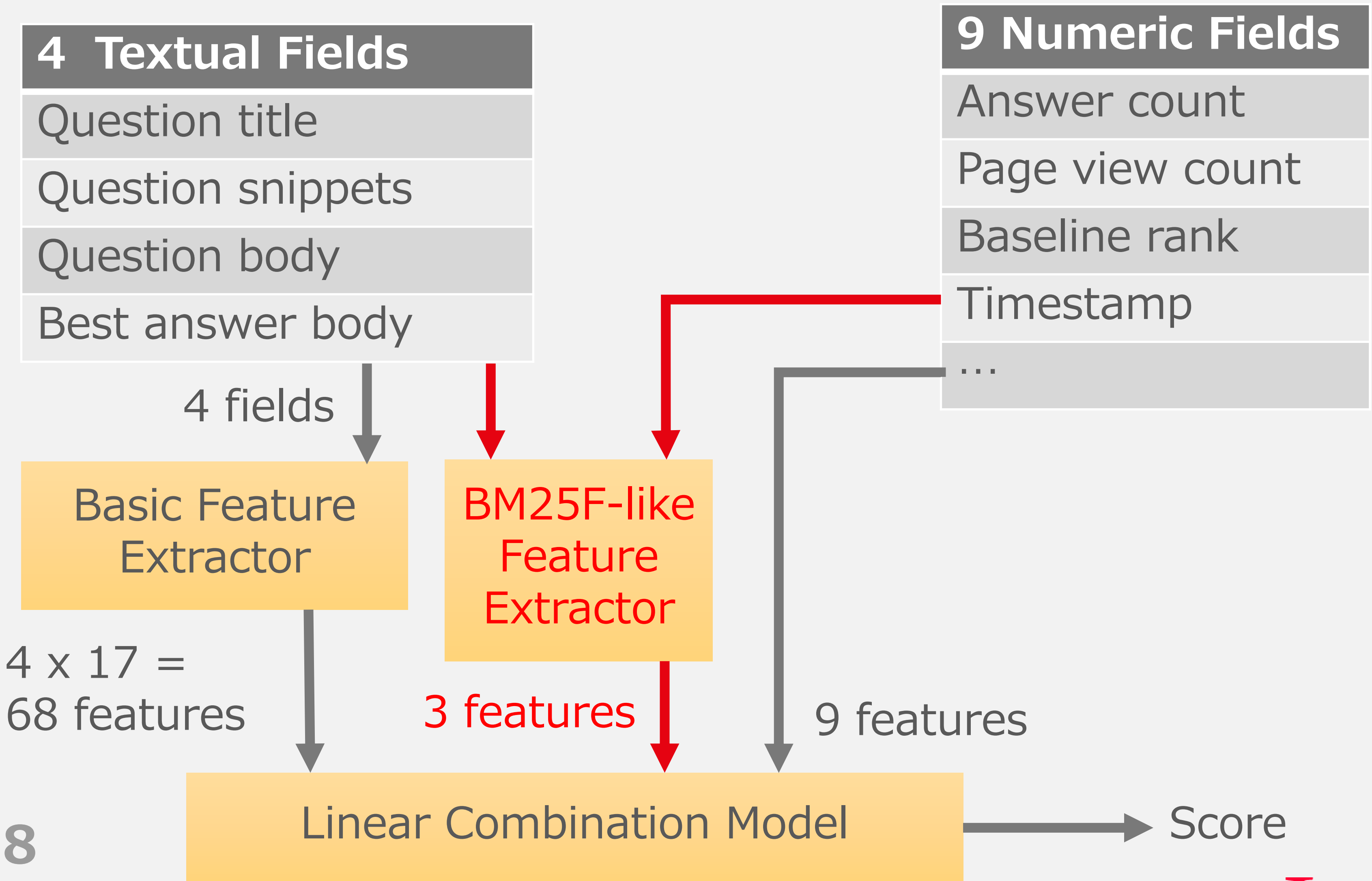
CA + BM25F

It is one of the best performers in the previous round of this task. It uses:

- 3 BM25F-like features in addition to the 77 basic features
- nDCG@10 as the objective function of CA
- 5-fold cross validation (5cv)

Ref:
Manabe *et al*. YJRS at the NTCIR-13 OpenLiveQ Task, *NTCIR* 2017.

CA + BM25F


4 Textual Fields
Question title
Question snippets
Question body
Best answer body
9 Numeric Fields
Answer count
Page view count
Baseline rank
Timestamp
…
4 fields
Basic Feature Extractor
BM25F-like Feature Extractor
4 x 17 =
68 features
3 features
9 features
Linear Combination Model
Score

CA + Translated

It is inspired from another best performer in the previous round.

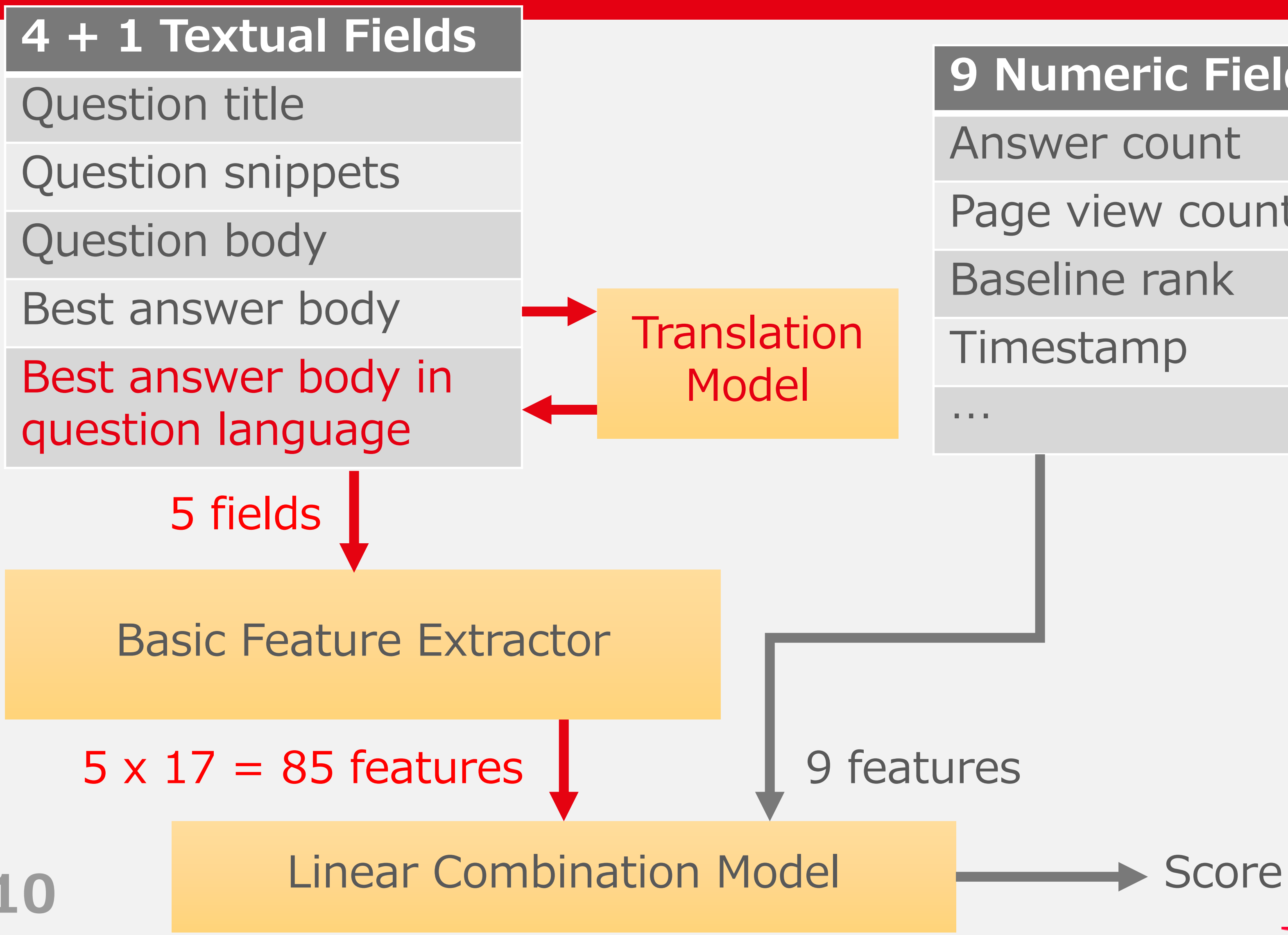
- Vocabulary of query must be more similar to questions than to answers.
- This method translates best answer body text into "*question language.*"
- The translation model is based on GIZA++ toolkit and a public corpus.

Refs: Chen *et al*. Erler at the NTCIR-13 OpenLiveQ task. NTCIR 2017

Och *et al*. Improved statistical alignment models. ACL 2000

Yahoo! *Chiebukuro* data (second edition)
https://www.nii.ac.jp/dsc/idr/yahoo/chiebkr2/Y_chiebukuro.html

CA + Translated

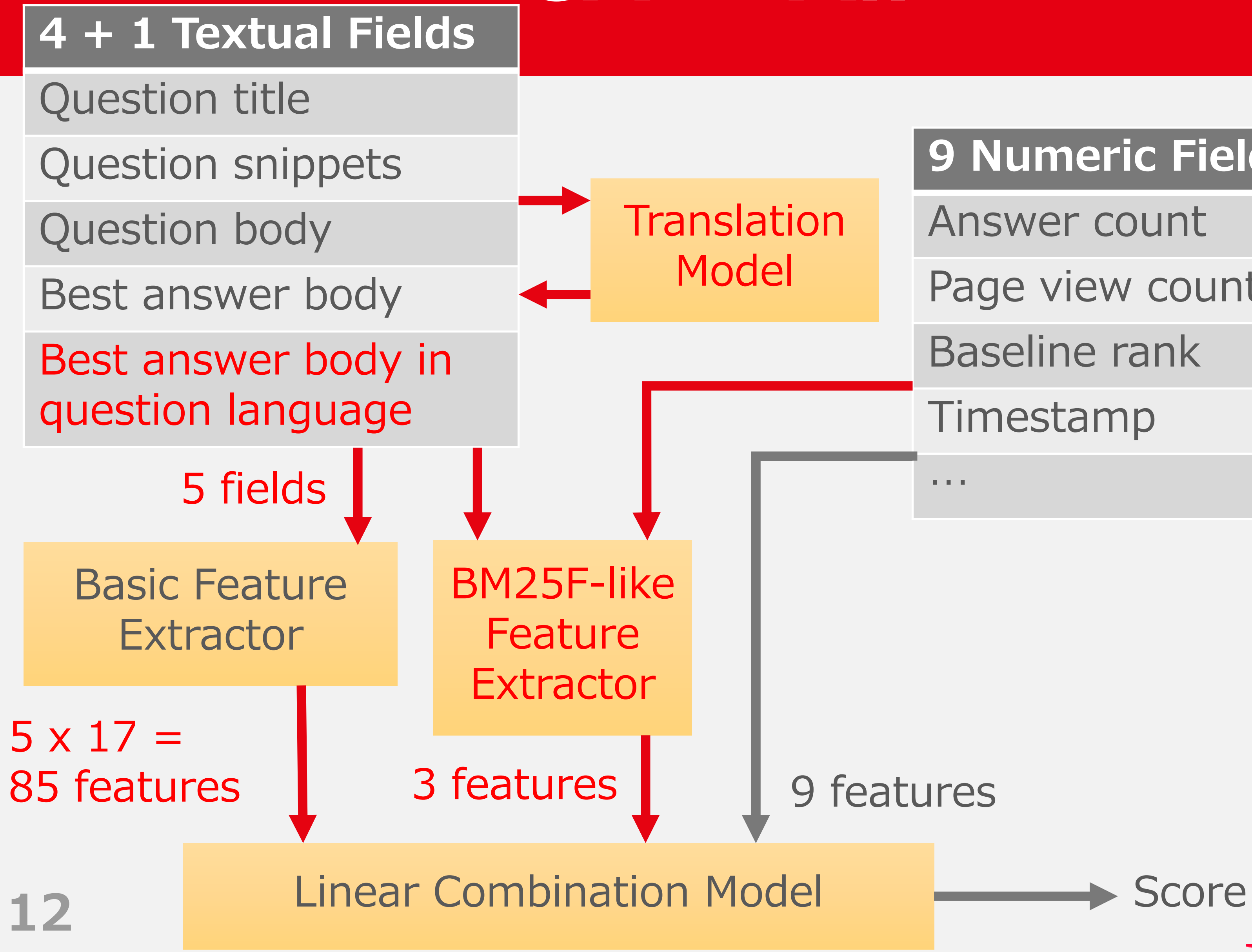


CA + All

It trains a model with all of:

- Basic features
- BM25F-like features
- Translated features

CA + All


4 + 1 Textual Fields
Question title
Question snippets
Question body
Best answer body
Best answer body in question language
Translation Model
9 Numeric Fields
Answer count
Page view count
Baseline rank
Timestamp
…
5 fields
Basic Feature Extractor
BM25F-like Feature Extractor
5 x 17 = 85 features
3 features
9 features
Linear Combination Model
Score

ListNet + Basic

We can arbitrary replace the linear combination model with a neural ranking model for example.

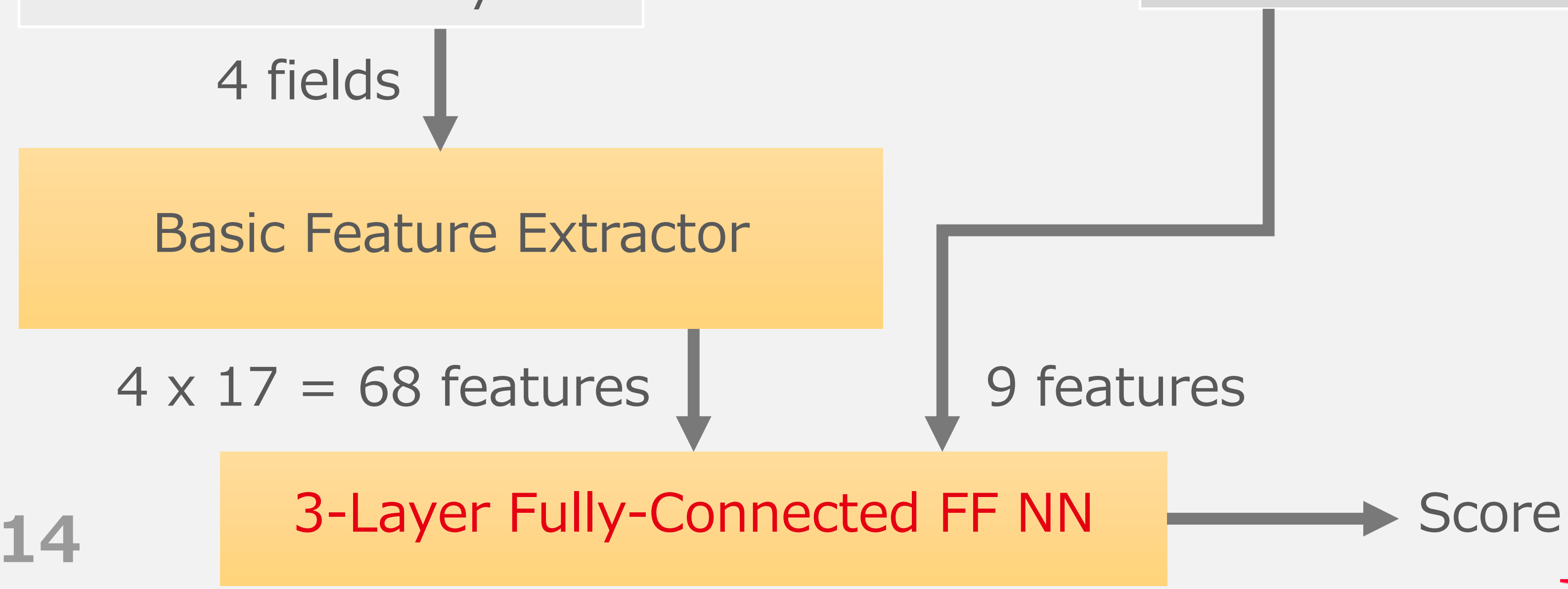
- We generated a simple neural ranking model with ListNet.
- The model is 3-layered fully-connected feed-forward neural network with 200 hidden nodes.
- It inputs 77 basic features.

Ref: Cao *et al*. Learning to rank:
from pairwise approach to listwise approach. ICML 2007

ListNet + Basic

4 Textual Fields
Question title
Question snippets
Question body
Best answer body

9 Numeric Fields
Answer count
Page view count
Baseline rank
Timestamp
…



LightGBM + Basic

Ensemble trees is another common model for learning to rank.

- We generated another simple model with LightGBM.
- It is linear combination of 100 regression trees with 15 leaves each.
- It inputs 77 basic features.

Ref: https://github.com/microsoft/LightGBM

LightGBM + Basic

4 Textual Fields
Question title
Question snippets
Question body
Best answer body

9 Numeric Fields
Answer count
Page view count
Baseline rank
Timestamp
…

4 fields

Basic Feature Extractor

4 x 17 = 68 features

9 features

Ensemble of 100 Regression Trees

Score

Minor Modifications

We also submitted slightly modified versions of our runs:

- A retry of CA+basic (ID: 95)
- ListNet+basic with 5cv (ID: 113)
- LightGBM+basic with a rough grid search on hyper-parameter space (ID: 148)

Note: Our methods probabilistically generate ranking models.

Other Possible Modifications

- We could input our extended feature sets into ListNet or LightGBM.
  - E.g., ListNet+BM25F, LightGBM+translated, …
- We did not do that due to the time limitation.

Summary of Our Runs
in descending order of Q-measure

Run ID	Description	Q-Measure
93	CA + translated	0.46387
92	CA + BM25F	0.45609
95	CA + basic (retry)	0.39559
91	CA + basic	0.39124
136	CA + all	0.38514
148	LightGBM + basic (tuned)	0.37429
100	ListNet + basic	0.37340
113	ListNet + basic (5cv)	0.37240
144	LightGBM + basic	0.37228

Evaluations

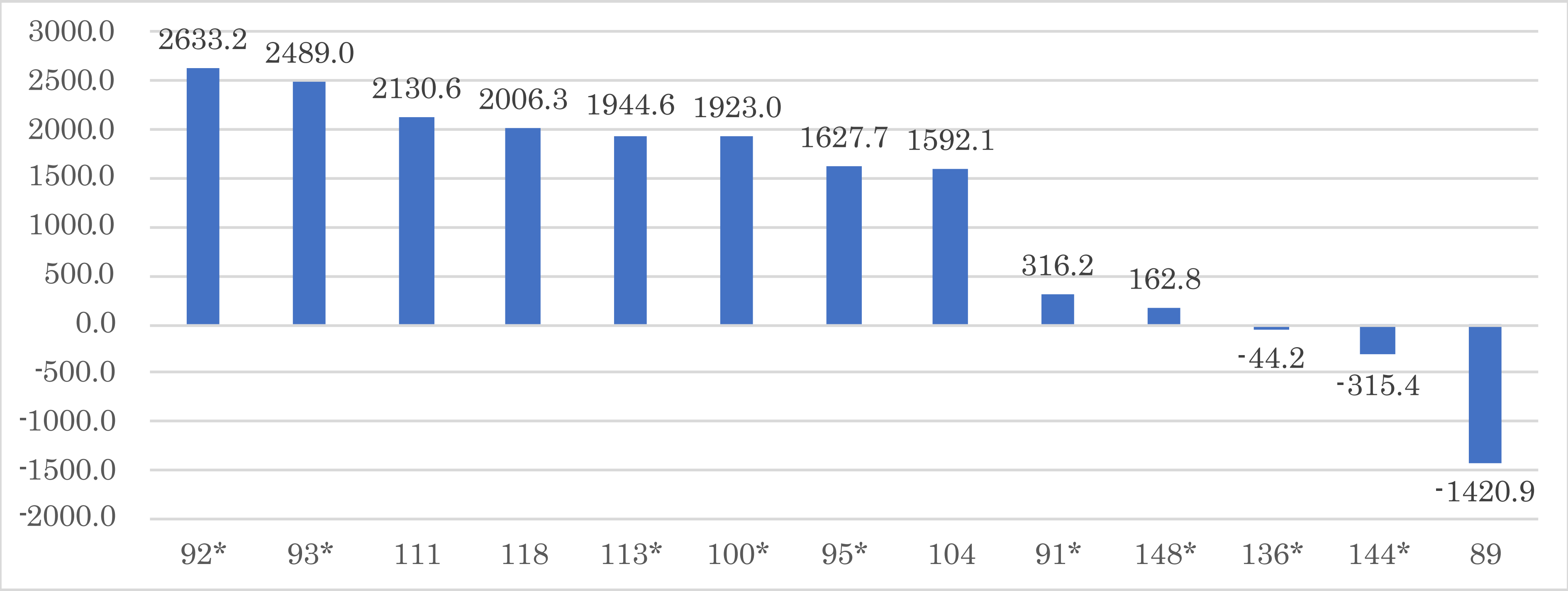
Evaluations

The evaluation process is the key point of this task!

It is based on Pairwise Preference Multileaving on Yahoo! *Chiebukuro.*

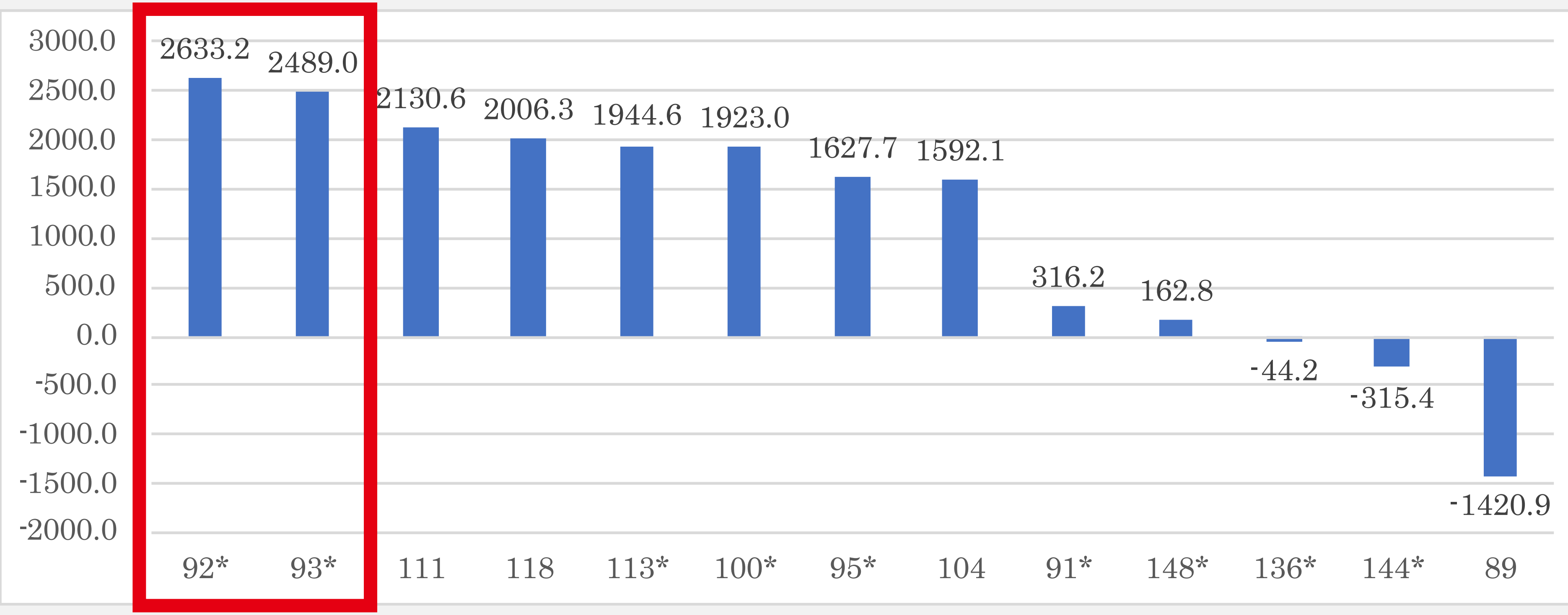
- First round: 61 runs (9 are ours)
- Second round: 30 runs (5 are ours)

First Round
credits of our runs* and other team's best runs



First Round

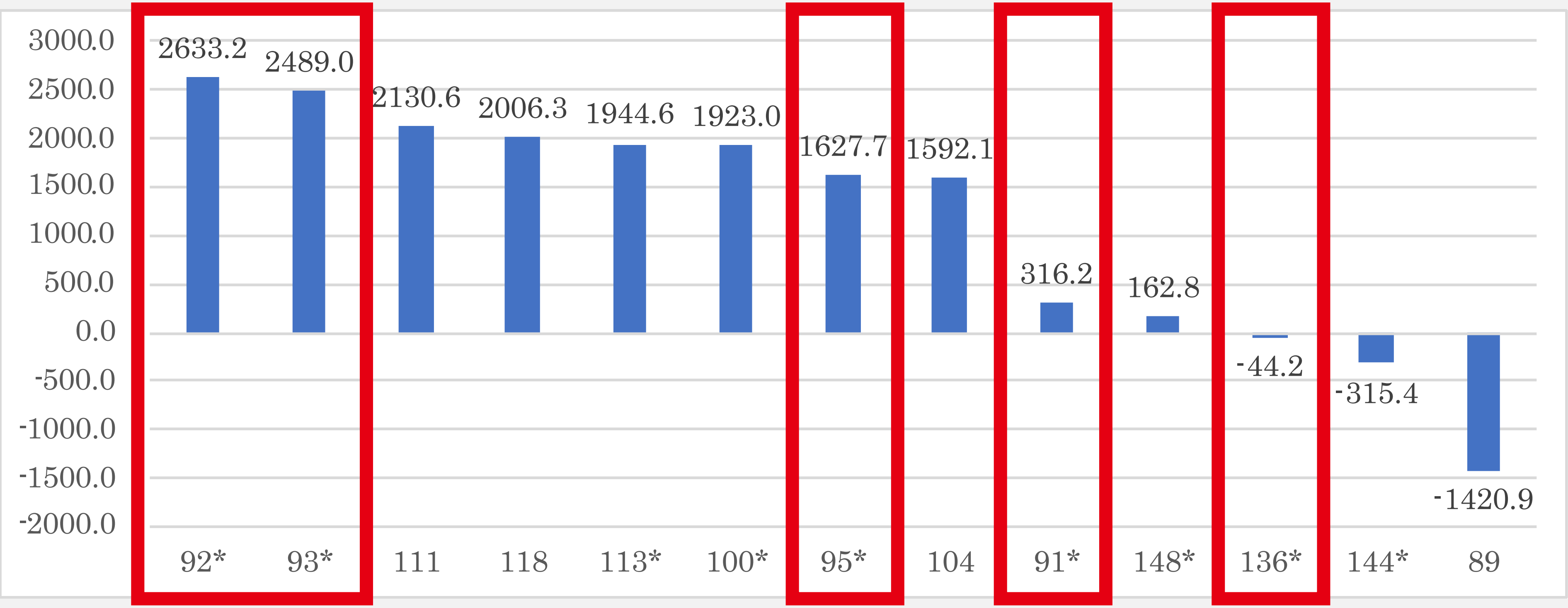
credits of our runs* and other team's best runs



Among our runs, CA+BM25F and CA+ translated runs were on the top-tier.
(consistent with offline evaluation)

First Round

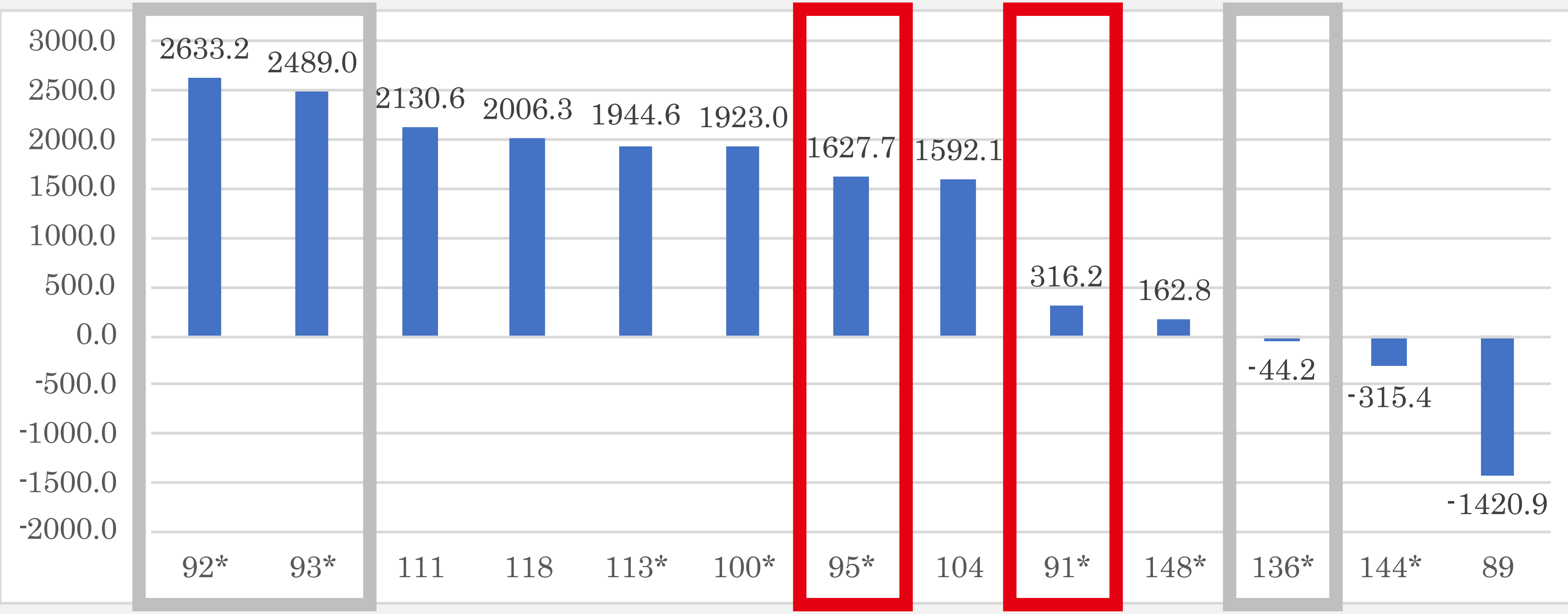
credits of our runs* and other team's best runs



The performance of our linear combination runs were not stable.

First Round

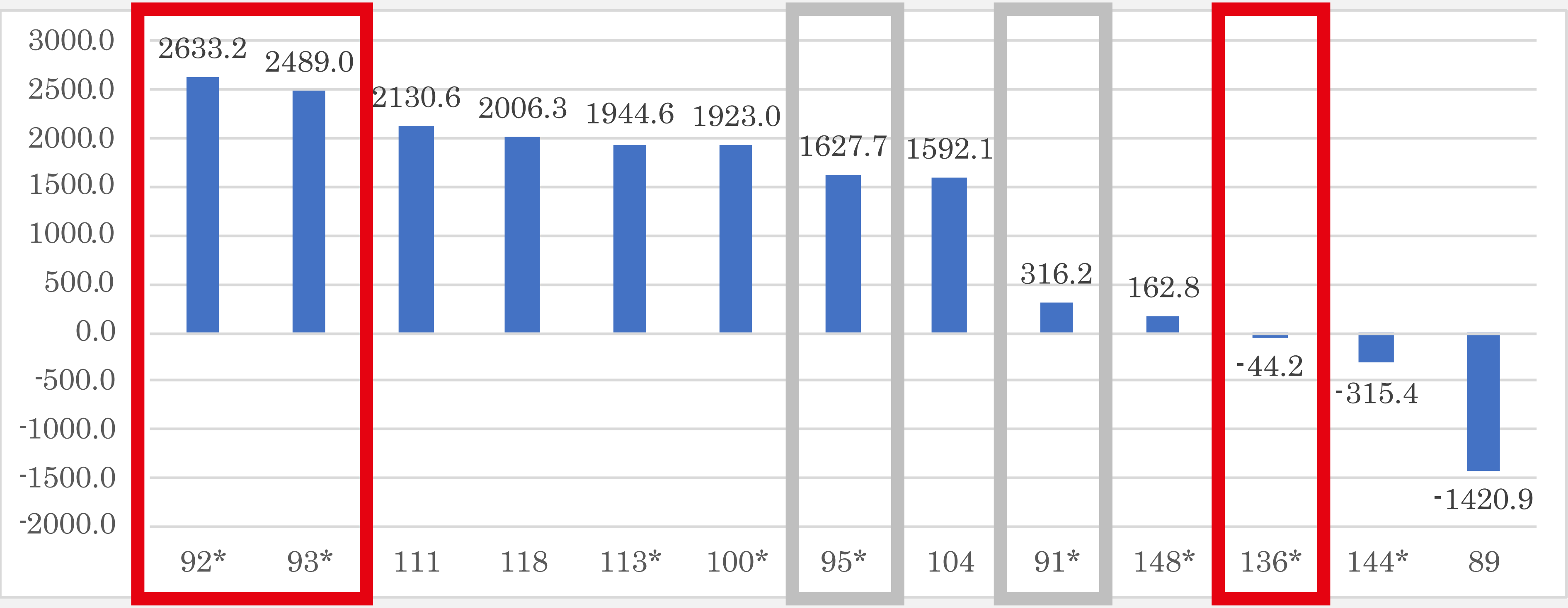
credits of our runs* and other team's best runs



As an example: Performance of runs 95 and 91 were quite different although they are both CA+basic.

First Round

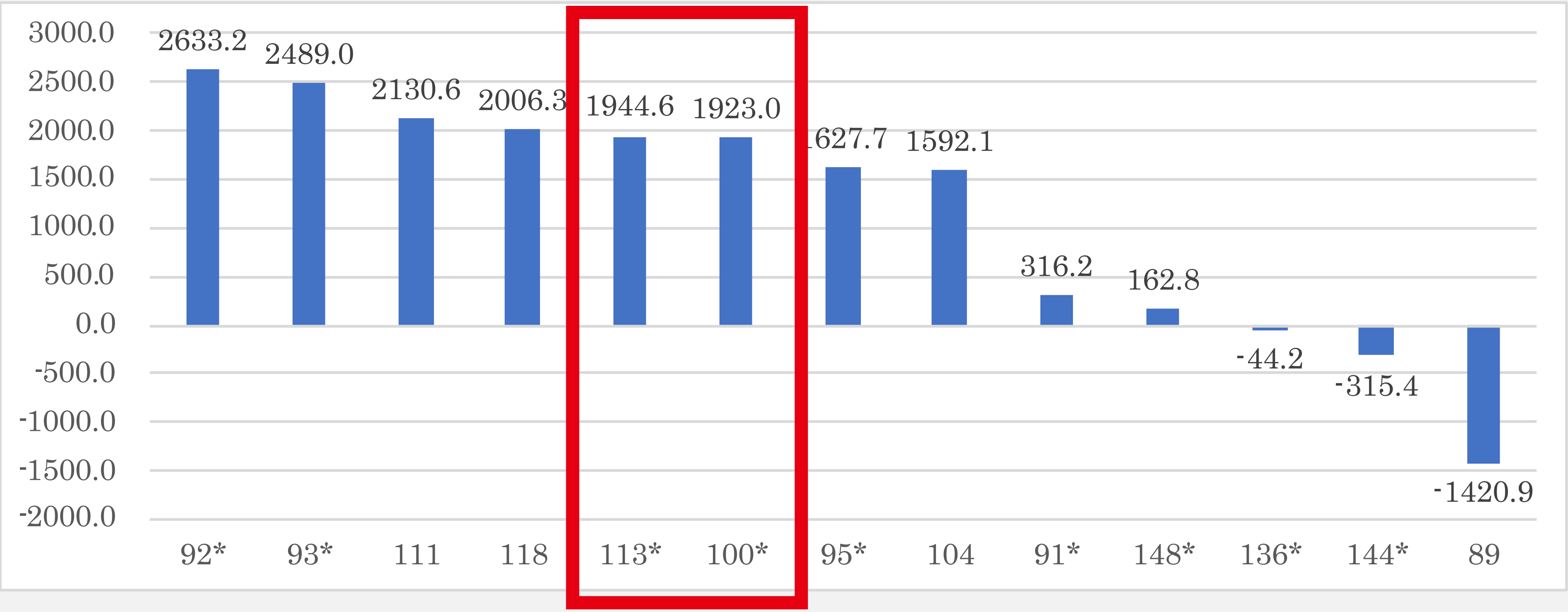
credits of our runs* and other team's best runs



Performance of CA+all was quite worse than CA+BM25F and CA+translated, which are the same configurations except using only subsets of all the features.

First Round

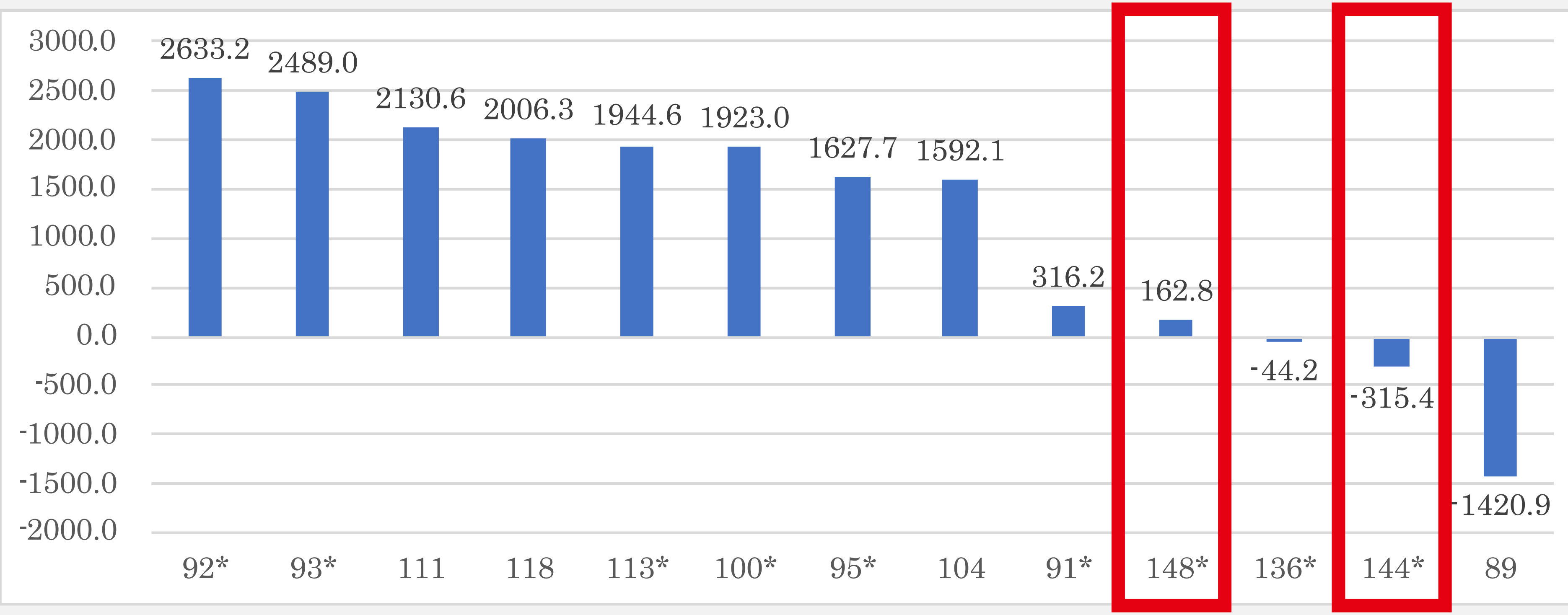
credits of our runs* and other team's best runs



Other than linear combination runs, ListNet+Basic runs were on the second-tier.

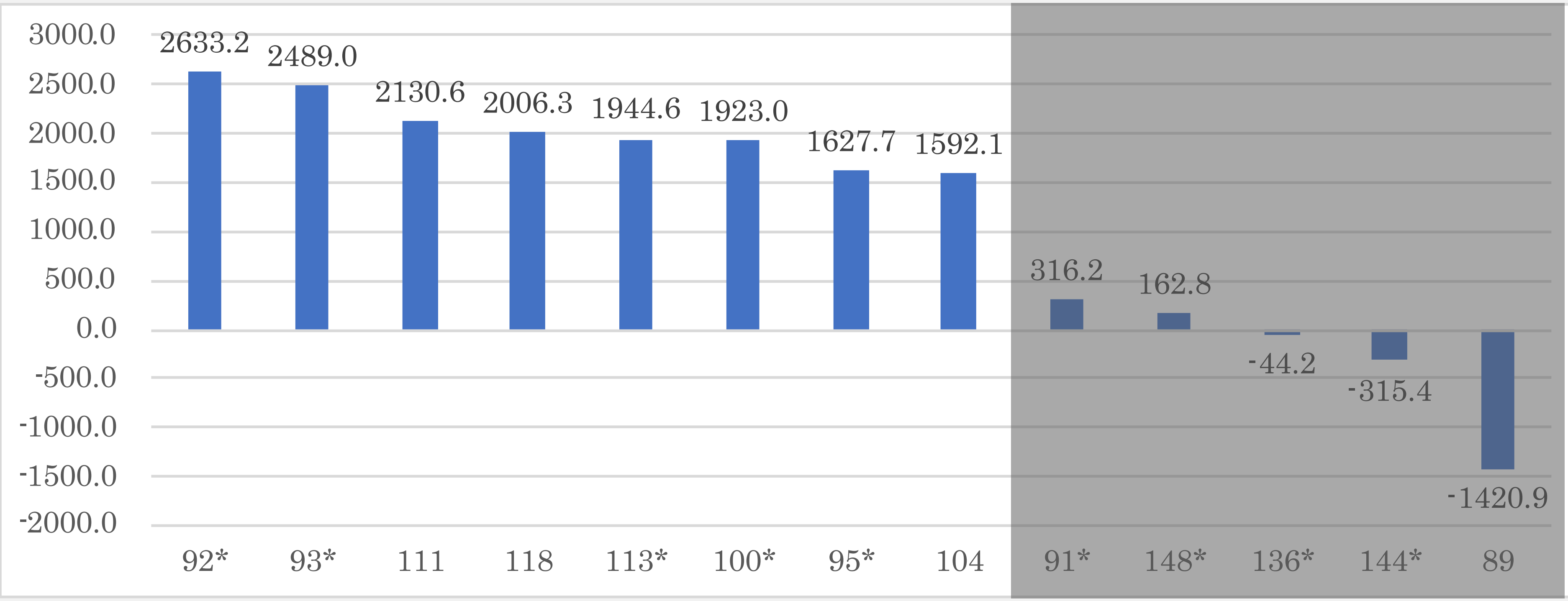
First Round

credits of our runs* and other team's best runs



Our LightGBM+Basic did not work well.

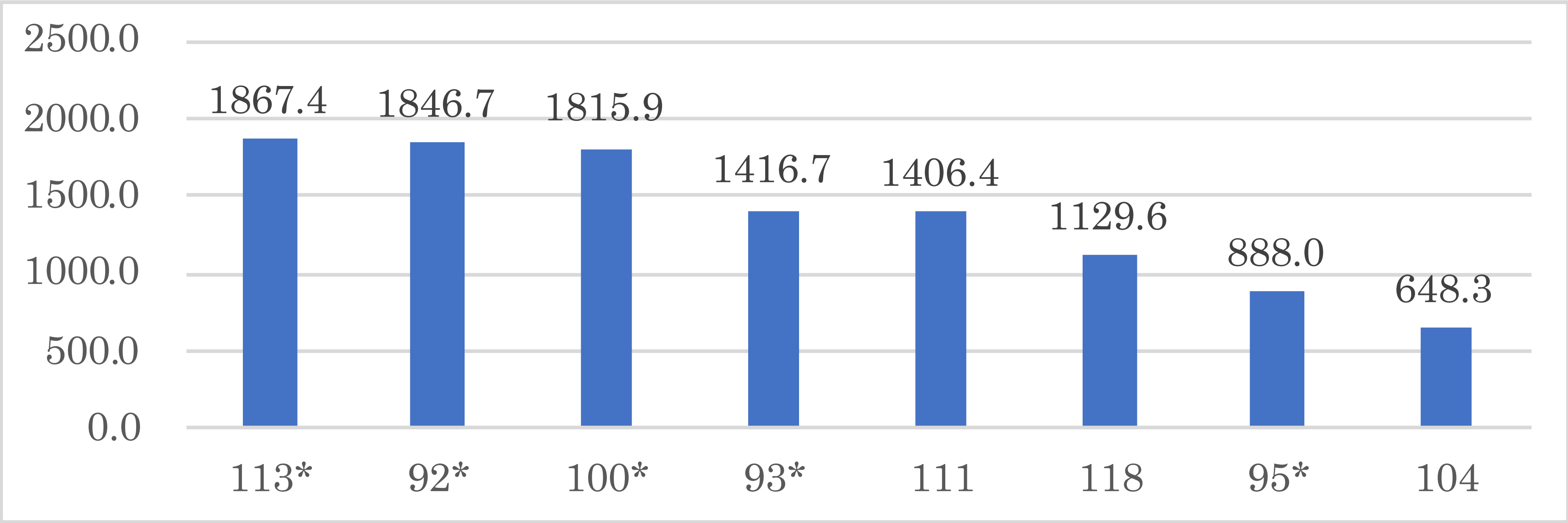
First Round
credits of our runs* and other team's best runs



Based on the result of this round, our LightGBM+basic and CA+all runs were excluded from the second round.

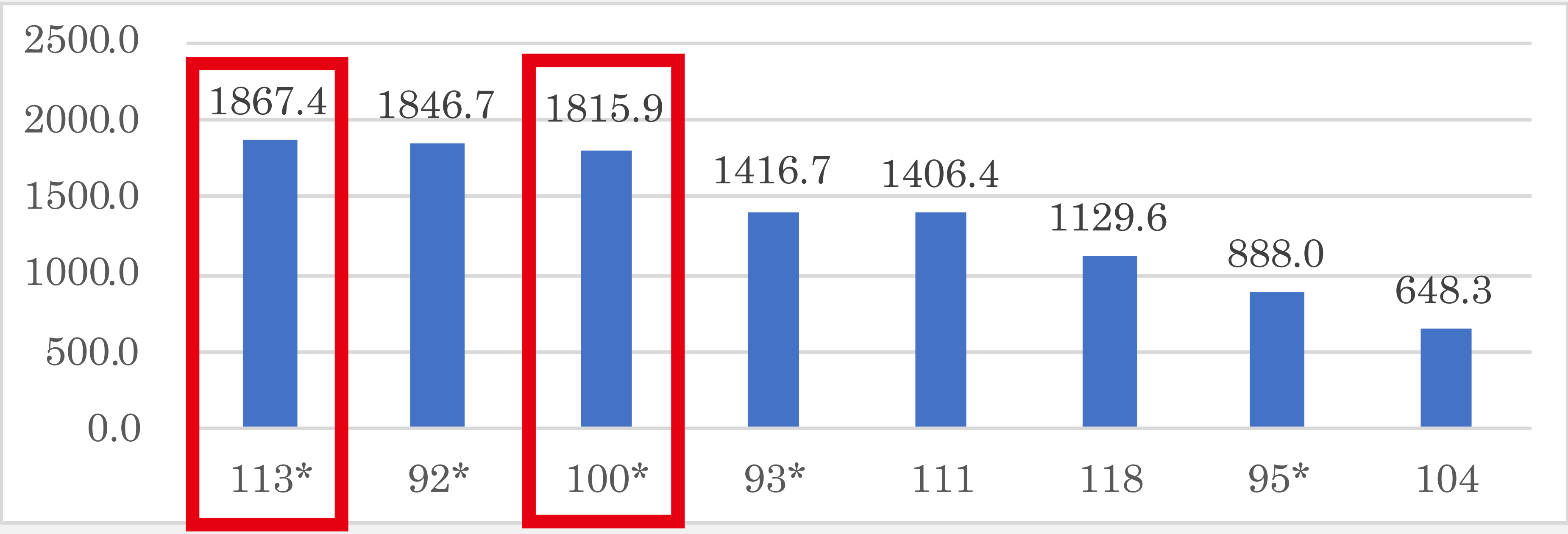
Second Round

credits of our runs* and other team’s best runs



Second Round

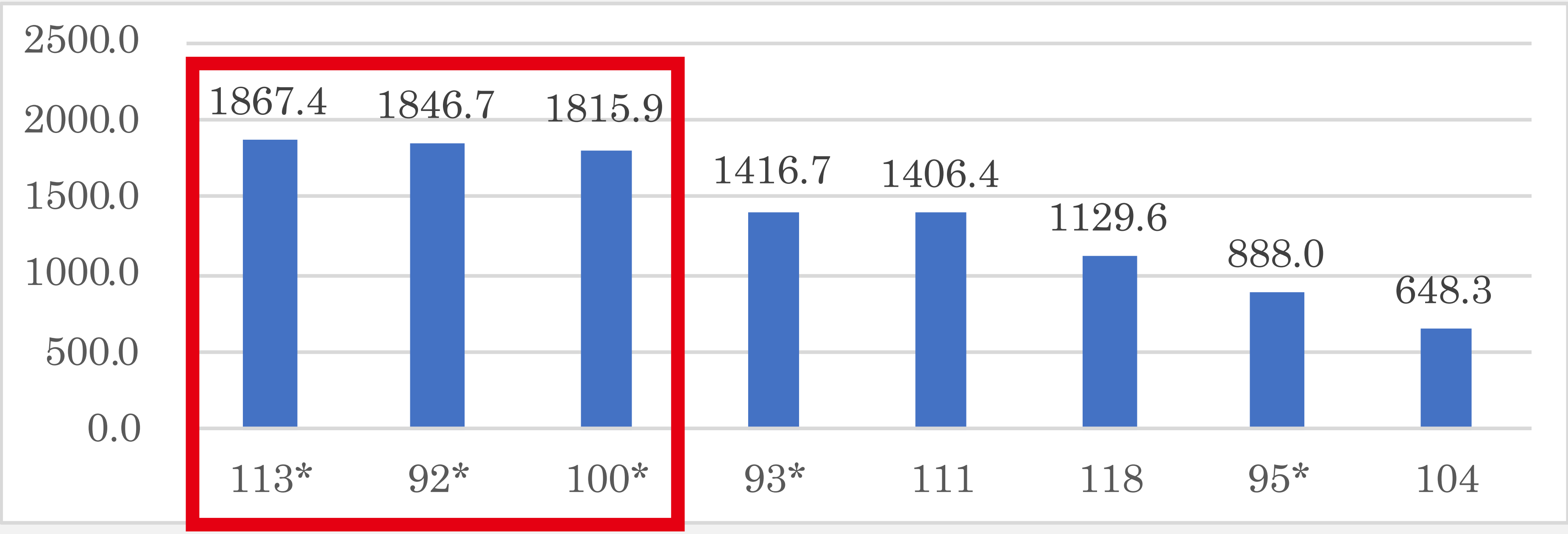
credits of our runs* and other team's best runs



Our ListNet+basic runs occupied better positions than in the first round.

Second Round

credits of our runs* and other team's best runs



The other tendencies were the same as in the first round.

Our ListNet+basic and CA+BM25F runs performed the best.

Conclusions

- Some combinations of a simple learning-to-rank method and reasonable features performed well.
- Namely, our ListNet+basic and CA+BM25F methods performed well.
- Our linear combinations resulted in unstable performance whereas ListNet+basic was quite promising.

EOP