A Systems and Setup: Details

For extractive systems, *K-Means* rank sentences clusters by descending order of cluster sizes, and then using a greedy algorithm (Lin and Bilmes, 2010) to select the nearest sentences to the centroid. Maximal Marginal Relevance (*MMR*) finds sentences which are highly relevant to the document but less redundant with sentences already selected for a summary. *cILP* (Gillick and Favre, 2009; Boudin et al., 2015) weights subsentences and maximizes their coverage by minimizing redundancy globally using Integer Linear Program (ILP). *TexRank* (Mihalcea and Tarau, 2004) automatically extracts keywords using Levenshtein distance between the text keywords. *LexRank* (Erkan and Radev, 2004) uses module centrality for ranking the keywords. In addition, we also use the recent three neural extractive systems: *CL* (Cheng and Lapata, 2016), *SumRun* (Nallapati et al., 2017), and *S2SExt* (Kedzie et al., 2018), where each has a little variation in their extraction architecture[13].

In training *CL*, *SumRun*, and *S2SExt*, we use upweight positive labels to make them proportional to the negative labels. We use 200 embedding size of GloVe (Pennington et al., 2014) pre-trained embeddings with 0.25 dropout on embeddings, fixing it not to be trained during training. We use CNN encoder with 6 window size as [25, 25, 50, 50, 50, 50] feature maps. We use 1-layer of sequence-to-sequence model with 300 size of LSTM and 100 size of MLP with 0.25 dropout. *SumRun* uses 16 size of segment and 16 size of position embeddings.

For abstractive systems, we use *WordILP* (Banerjee et al., 2015) that produces a word graph of important sentences and then choose sentences from the word graph employing a ILP solver. We also use incremental sequence-to-sequence models: a basic *S2SAbs* (Rush et al., 2015) with *Pointer* network (See et al., 2017), with teacher forcing *Teacher* (Bengio et al., 2015), and with reinforcement learning on the evaluation metrics, and *RL* (Paulus et al., 2017).

In training *S2SAbs*, *Pointer*, *Pointer*, and *RL*, we use 150 hidden size of GRU with 300 size of GloVe embeddings. *Pointer* uses maximum coverage function using NLL loss. *Teacher* uses 0.75 ratio of teach forcing with exponential decaying function. and *RL* uses 0.1 ratio of RL optimization after the first epoch of *S2SAbs* training. We use 4 size of beam searching at decoding. We use 32 batch size with adam optimizer of 0.001 learning rate.

For `MScript`, the original dataset has no data split, so we randomly split it by 0.9, 0.05, 0.05 for train, valid, test set, respectively.

[13] See (Kedzie et al., 2018) for a detailed comparison.

B Venn Diagram for All Datasets

Sentence Venn diagrams among three aspects and oracle for all datasets are shown in Figure 6. `Newsroom` has an analogous pattern to `XSum`. Compared to `PeerRead`, , `PubMed` has relatively less sentence overlap between First-k and the other two aspects. `MScript` has extremely small oracle sentence overlaps to all three aspects. However, it is mainly because of the characteristics of the dataset: it has long source documents (1k sentences on average) with short (5 sentences on average) summary.

C Full ROUGE F Scores for Corpus Bias Analysis

In Table 5, we provide a full list of ROUGE F scores for all datasets w.r.t three sub-aspects. We find that in `MScript`, the best algorithms for each of ROUGE-1/2/L are different.

D Documents in an Embedding Space: for All Datasets

In Figure (7,8), we have more two-dimensional PCA projection examples for source documents from all datasets. We find a weak pattern about where target sentences lie on according to the number of them. For example, from `XSum` and `Reddit` which have a single target sentence, we investigate that some target sentences are located in the middle of `ConvexHull`, which are far from any source sentences.

E System Biases per each corpus with the Three Sub-aspects

In Figure 9, we have more diagrams showing system biases toward each of three sub aspects. We find that there exists a bias according to the corpus: for example in `Reddit`, many systems have a importance bias in common. On the other hand, systems are biased toward a diversity aspect in `AMI`. Also, some systems tend to be biased in certain aspect across the different corpus: systems such as

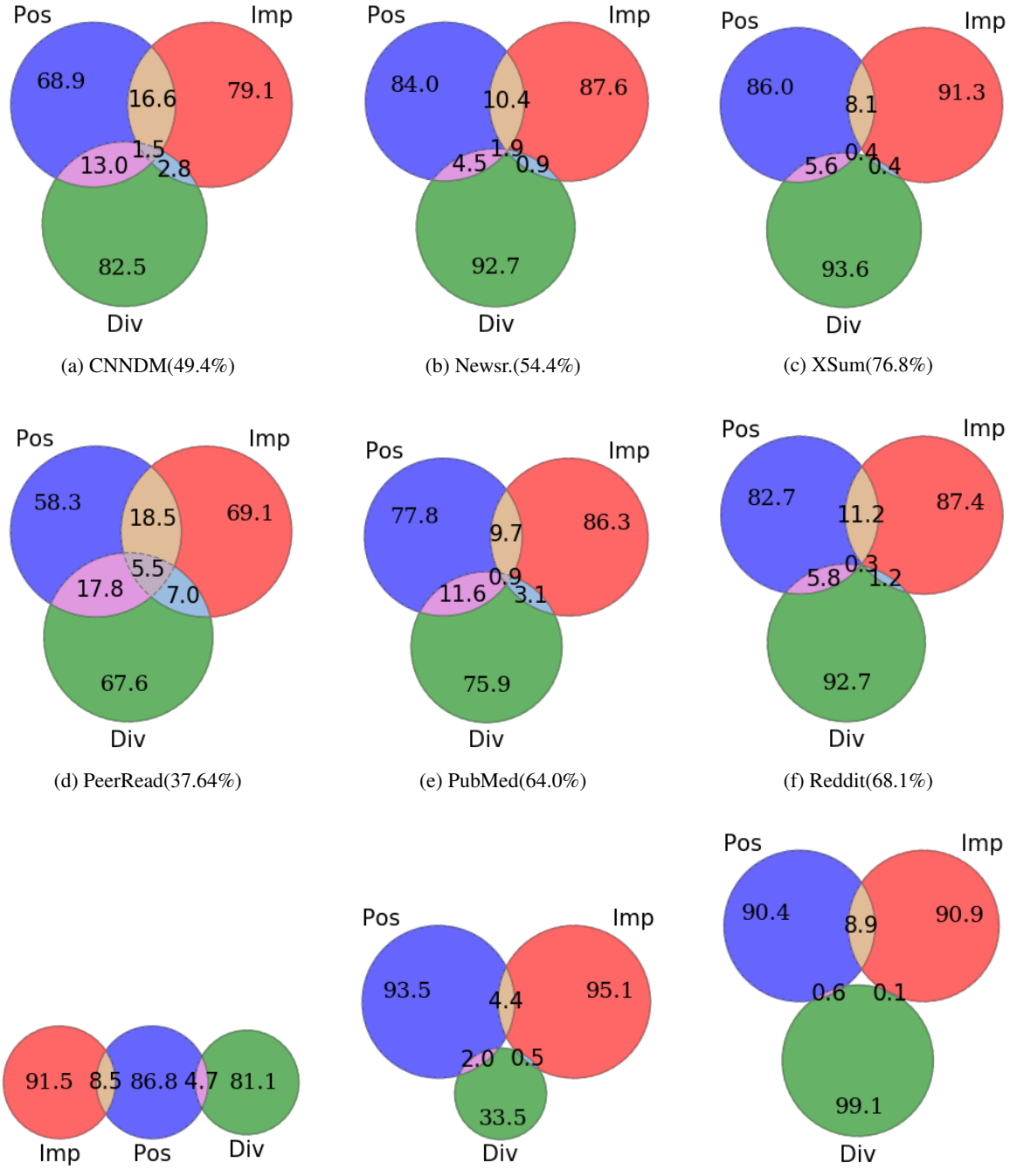
KMeans and *MMR*, many corpora are biased toward a importance aspect.

		CNNDM	NewsRoom	XSum
		R-1/2/L	R-1/2/L	R-1/2/L
	Random	26.6/6.7/23.9	15.2/2.8/12.2	14.9/1.8/11.2
	Oracle	51.5/28.5/48.6	53.4/40.2/50.7	27.9/7.5/23.2
Position	First-k	**39.1/17.1/35.8**	**36.9/25.9/33.9**	14.8/1.4/11.1
Position	Last-k	23.5/4.7/21.1	11.5/2.0/9.5	13.2/1.5/10.1
Position	Middle-k	29.4/8.6/26.4	17.4/5.3/14.4	14.7/1.7/11.0
Divers.	ConvexFall	29.5/8.6/26.6	15.0/4.0/12.7	13.6/1.3/10.5
Divers.	Heuristic	29.2/8.7/26.3	14.9/4.1/12.7	13.6/1.3/10.5
Import.	N-Nearest	29.7/9.3/26.9	18.9/6.1/15.7	**15.7/2.0/11.7**
Import.	K-Nearest	30.6/10.5/27.8	19.1/6.8/16.0	15.0/1.8/11

		PeerRead	PubMed	Reddit
		R-1/2/L	R-1/2/L	R-1/2/L
	Random	38.2/11.1/34.3	41.3/11.3/37.6	17.6/3.7/14.2
	Oracle	56.6/29.5/52.7	58.2/27.9/54.8	38.5/17.8/33.8
Position	First-k	**41.4/16.8/37.9**	37.8/10.2/34.7	**21.8/6.2/17.8**
Position	Last-k	39.1/12.4/35.1	39.1/11.8/35.9	116.4/3.7/13.4
Position	Middle-k	40.4/12.5/36.3	39.5/10.8/36.3	17.4/3.2/13.8
Divers.	ConvexFall	40.4/12.8/36.3	39.0/10.3/35.3	17.3/3.2/14.2
Divers.	Heuristic	39.7/12.4/35.6	38.1/9.8/34.5	17.2/3.2/14.2
Import.	N-Nearest	**41.4**/13.2/37.3	**43.1/12.7/39.5**	20.6/4.4/16.5
Import.	K-Nearest	41.0/14.0/36.9	40.0/12.3/36.6	15.1/3.6/12.3

		AMI	BookSum	MScript
		R-1/2/L	R-1/2/L	R-1/2/L
	Random	17.4/2.2/16.3	41.6/7.0/39.6	12.2/0.7/11.3
	Oracle	42.8/12.3/40.9	52.0/14.7/50.2	33.5/7.3/31.7
Position	First-k	16.4/2.3/15.5	**40.8/7.6/38.9**	10.3/**1.1**/9.4
Position	Last-k	11.1/1.7/10.5	37.6/5.8/36.1	**13.4**/0.9/12.1
Position	Middle-k	16.1/1.9/15.2	39.4/6.6/37.7	12.1/0.6/11.2
Divers.	ConvexFall	**20.4/2.5/19.1**	24.3/3.9/22.6	12.8/0.7/11.9
Divers.	Heuristic	15.7/1.5/15.0	38.2/6.2/36.4	9.7/0.5/9.1
Import.	N-Nearest	1.9/0.1/1.8	39.3/6.9/37.4	13.1/0.8/**12.2**
Import.	K-Nearest	0.0/0.0/0.0	30.9/5.0/29.5	1.0/0.0/1.0

Table 5: Full ROUGE-1/2/L F-Scores for different corpora w.r.t three sub-aspects algorithms.

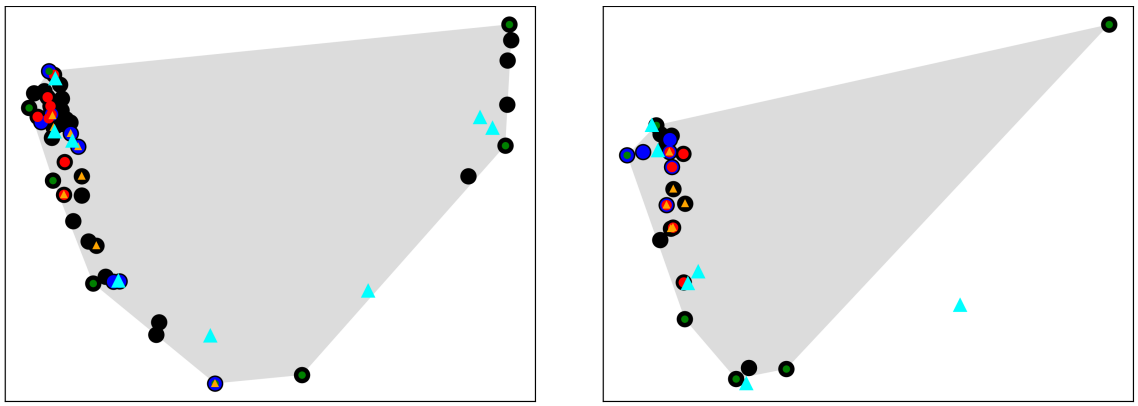


(a) CNNDM(49.4%) (b) Newsr.(54.4%) (c) XSum(76.8%)

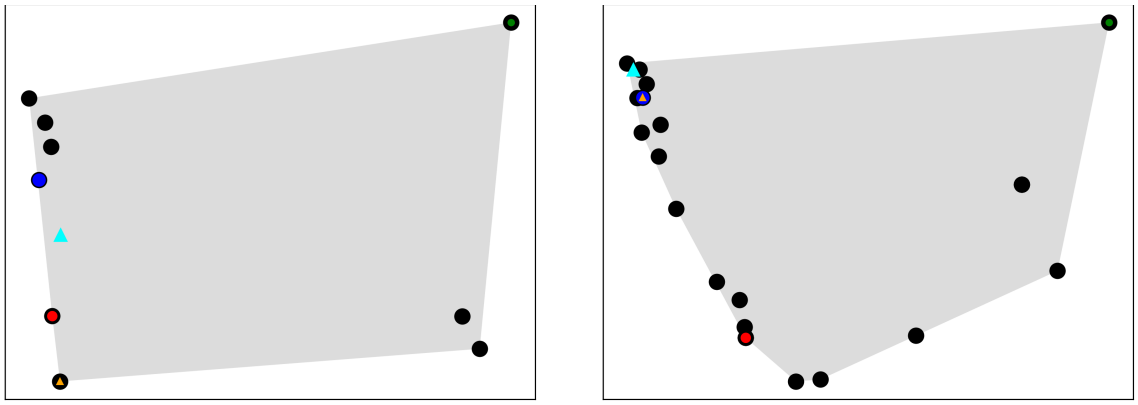
(d) PeerRead(37.64%) (e) PubMed(64.0%) (f) Reddit(68.1%)

(g) AMI(94.1%) (h) BookSum(87.1%) (i) MScript (99.1%)

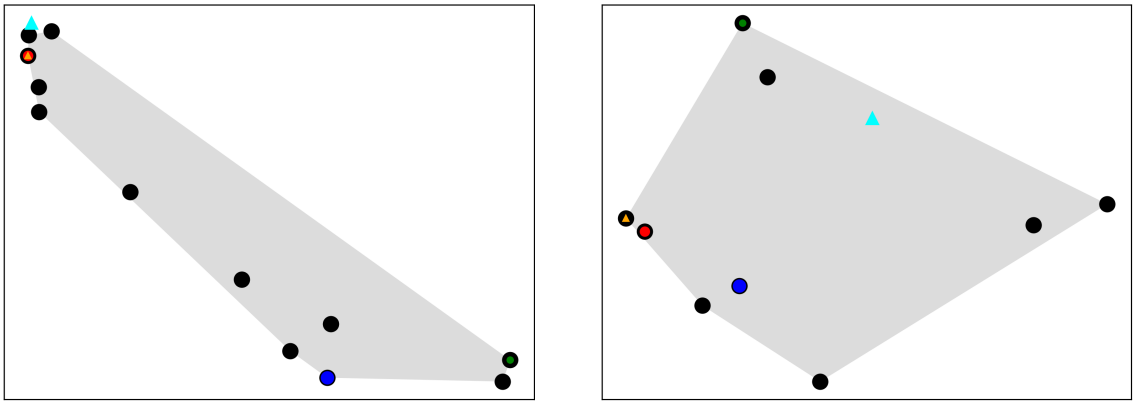
Figure 6: Venndiagram of averaged summary sentence overlaps across the the sub-aspects for all datasets. We use First-k for Position (P), ConvexFall for Diversity (D), and N-Nearest for Importance (I). The number called *Oracle Recall* in the parenthesis is the averaged ratio of how many the oracle sentences are NOT chosen by union set of the three sub-aspect algorithms.



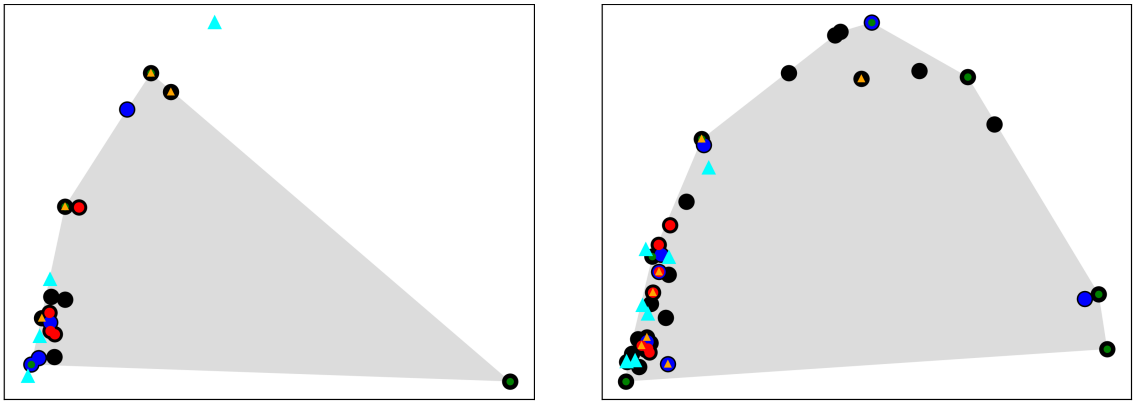
(a) CNNDM



(b) NewsRoom

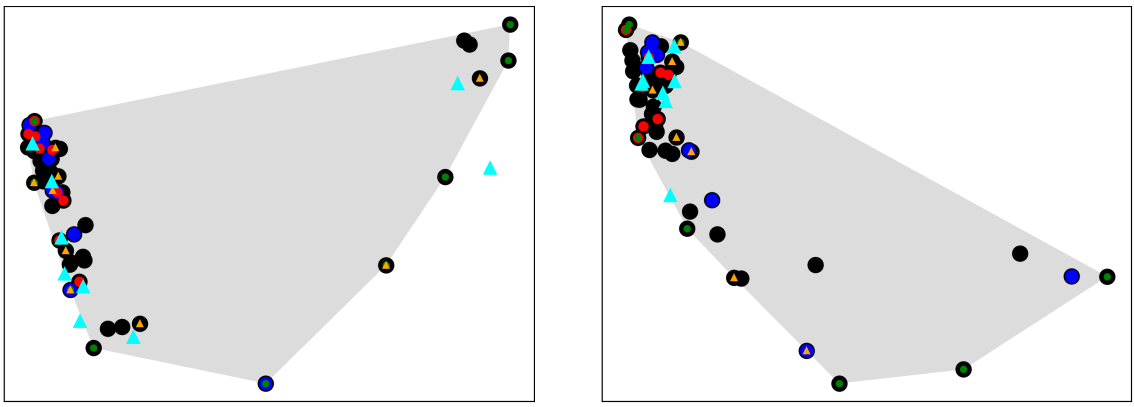


(c) XSum

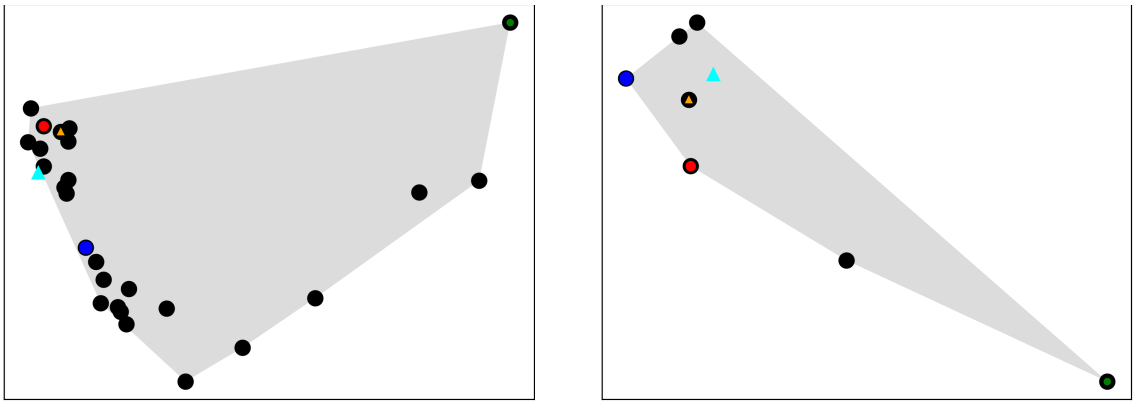


(d) PeerRead

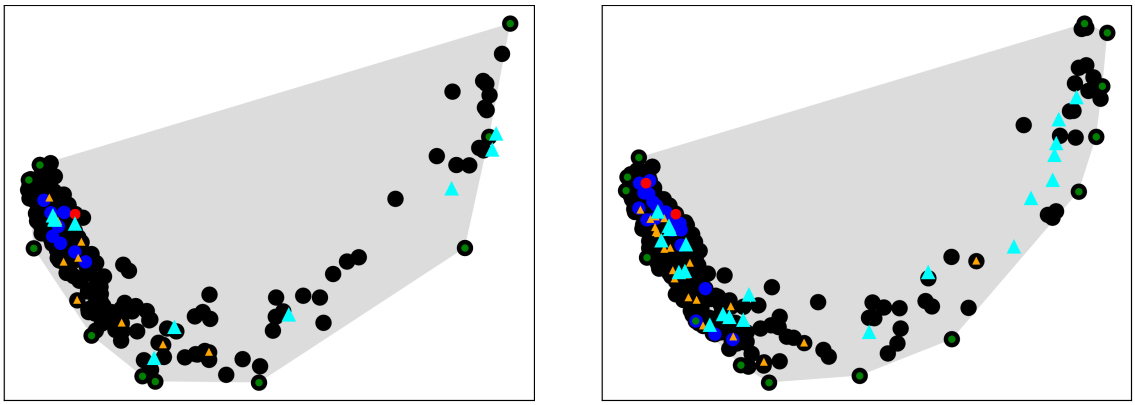
Figure 7: PCA projection of extractive summaries chosen by multiple aspects of algorithms (`CNNDM`, `NewsRoom`, `XSum`, `PeerRead`, and `PubMed`). Source and target sentences are black circles (●) and purple stars, respectively. The blue, green, red circles are summary sentences chosen by `First`, `ConvexFall`, `KN`, respectively. The yellow stars are the oracle sentences. Best viewed in color.



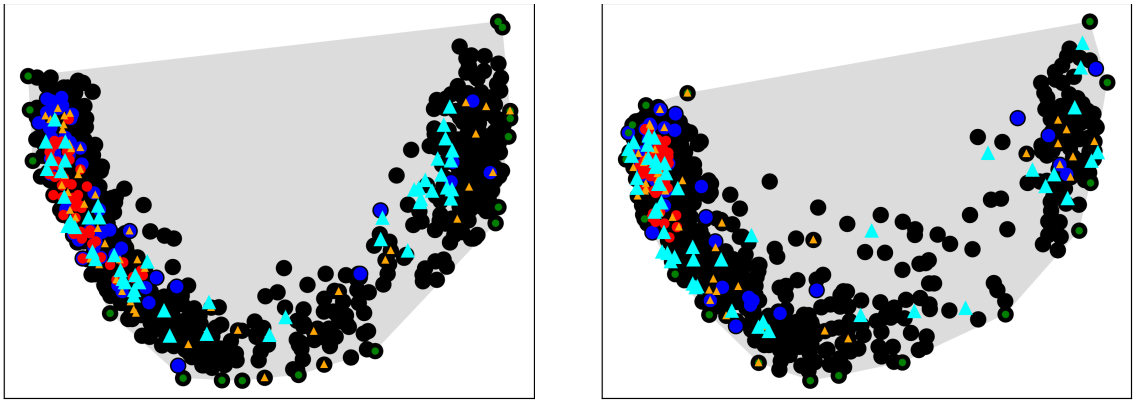
(a) PubMed



(b) Reddit

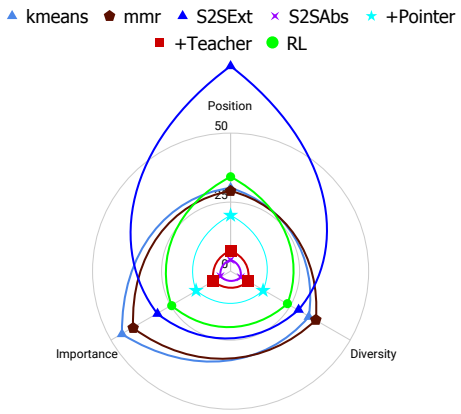


(c) AMI

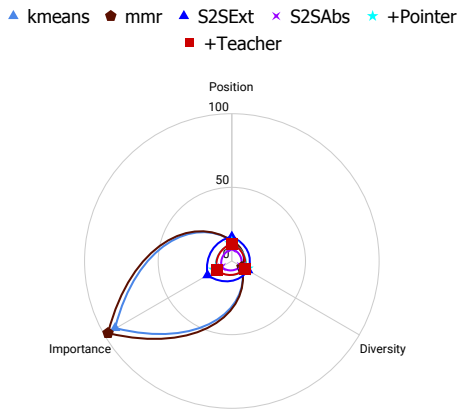


(d) BookSum

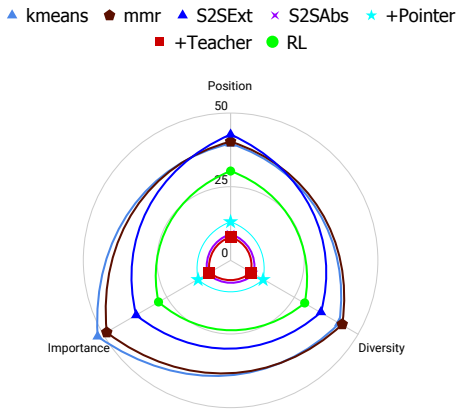
Figure 8: PCA projection of extractive summaries chosen by multiple aspects of algorithms (`Reddit`, `AMI`, `Booksum`, and `MScript`). Source and target sentences are black circles (●) and purple stars, respectively. The blue, green, red circles are summary sentences chosen by `First`, `ConvexFall`, `KN`, respectively. The yellow stars are the oracle sentences. Best viewed in color.



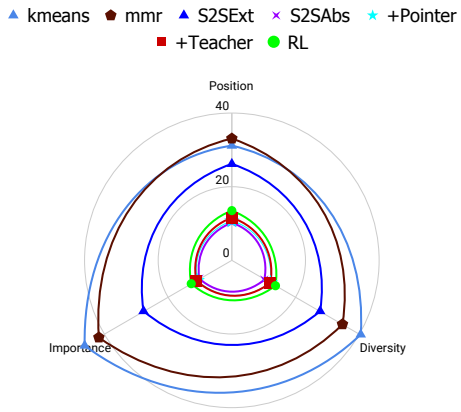
(a) CNNDM



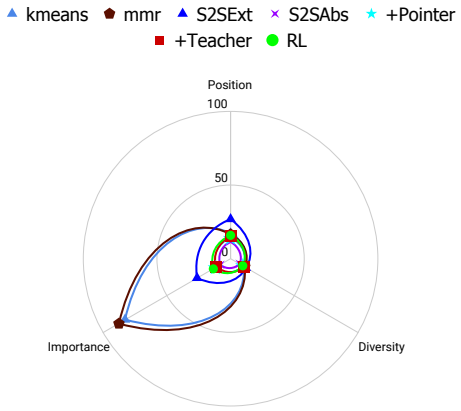
(b) XSum



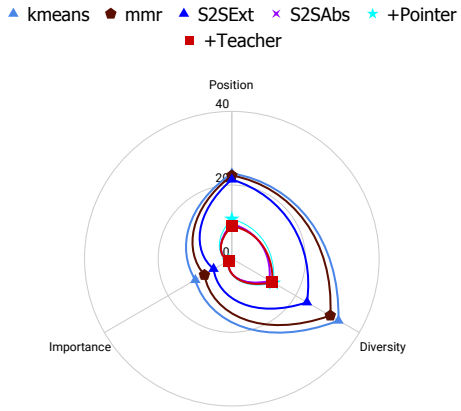
(c) PeerRead



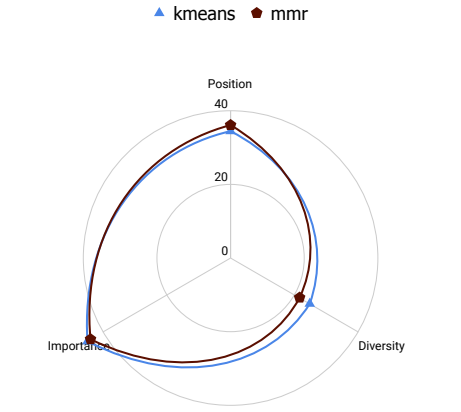
(d) PubMed



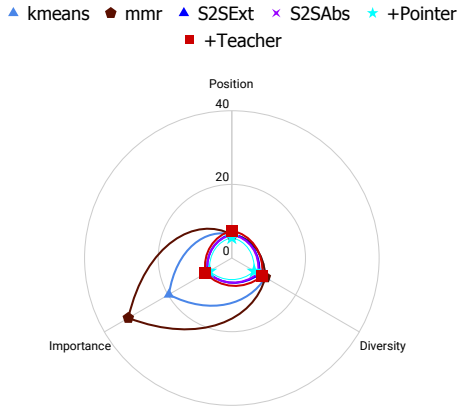
(e) Reddit



(f) AMI



(g) BookSum



(h) MScript

Figure 9: System biases with the three sub-aspects per each corpus, showing what portion of aspect is used for each system.