Datasets	ICSI			UB			ACCEPT			JOINT			AL			AL+		
	R1	R2	SU4	R1	R2	SU4	R1	R2	SU4	R1	R2	SU4	R1	R2	SU4	R1	R2	SU4
Concept Notion: Bigrams																		
DBS	.451	.183	.196	.848	.756	.532	.778	.654	.453	.815	.707	.484	**.833**	**.729**	.498	.828	.721	**.500**
DUC'04	.374	.090	.118	.470	.212	.185	.442	.176	.165	**.444**	**.180**	**.166**	.440	.178	.160	.427	.166	.154
DUC'02	.350	.085	.110	.474	.216	.187	.439	.178	.161	.444	.182	.165	**.448**	**.188**	.165	**.448**	.184	**.170**
DUC'01	.333	.073	.105	.450	.213	.181	.414	.171	.156	.418	.167	.149	**.435**	**.186**	**.163**	.426	.181	.158
Concept Notion: Content Phrases																		
DBS	.403	.135	.154	.848	.756	.532	.691	.581	.436	.742	.597	.419	**.776**	**.652**	**.448**	.767	.629	.440
DUC'04	.374	.090	.118	.470	.212	.185	.441	.176	.160	.441	.179	.162	**.444**	**.180**	**.162**	.422	.164	.150
DUC'02	.350	.085	.110	.474	.216	.187	.436	.181	.162	.444	.183	.165	**.446**	**.185**	**.168**	.442	.182	.162
DUC'01	.333	.073	.105	.450	.213	.181	.410	.165	.153	.417	.170	.156	**.433**	**.182**	**.161**	.420	.179	.154

Table 3: ROUGE-1 (R1), ROUGE-2 (R2) and ROUGE SU-4 (SU4) achieved by our models after the tenth iteration of the interactive loop in comparison to the upper bound and the basic ILP setup

Datasets	ACCEPT #F	JOINT #F	AL #F	AL+ #F
Concept Notion: Bigrams				
DBS	313	296	348	342
DUC'04	15	14	16	14
DUC'02	14	13	15	15
DUC'01	13	11	13	13
Concept Notion: Content Phrases				
DBS	110	114	133	145
DUC'04	8	9	10	10
DUC'02	7	7	8	6
DUC'01	7	7	8	6

Table 4: Average amount of user feedback (#F) considered by our models at the end of the tenth iteration of the interactive summarization loop

the summary for an individual user, we evaluate our models based on the mean ROUGE scores across clusters per reference summary. In Table 4, we additionally evaluate the models based on the amount of feedback (#F $= |I_0^T|$) taken by the oracles to converge to the upper bound within ten iterations.

To examine the system performance based on user feedback, we analyze our models' performance on multiple datasets. The results in Table 3 show that our idea of interactive multi-document summarization allows users to steer a general summary towards a personalized summary consistently across all datasets. From the results, we can see that the AL model starts from the concept-based ILP summarization and nearly reaches the upper bound for all the datasets within ten iterations. AL+ performs similar to AL in terms of ROUGE, but requires less feedback (compare Table 4). Furthermore, the ACCEPT and JOINT models get stuck in a local optimum due to the less exploratory nature of the models.

5.2 Concept Notion

Our interactive summarization approach is based on the scalable global concept-based model which uses bigrams as concepts. Thus, it is intuitive to use bigrams for collecting user feedback as well.[7] Although our models reach the upper bound when using bigram-based feedback, they require a significantly large number of iterations and much feedback to converge, as shown in Table 4.

To reduce the amount of feedback, we also consider content phrases to collect feedback. That is, syntactic chunks from the constituency parse trees consisting of non-function words (i.e., nouns, verbs, adjectives, and adverbs). For DBS being extractive dataset, we use bigrams and content phrases as concepts, both for the objective function in equation (1) and as feedback items, whereas for the DUC datasets, the concepts are always bigrams for both the feedback types (bigrams/content phrases). For DUC being abstractive, in the case of feedback given on content phrases, they are projected back to the bigrams to change the concept weights in order to have more overlap of simulated feedback. Table 4 shows feedbacks based on the content phrases reduces the number of feedbacks by a factor of 2. Furthermore, when content phrases are used as concepts for DBS, the performance of the models is lower compared to bigrams, as seen in Table 3.

5.3 Datasets

Figure 2 compares the ROUGE-2 scores and the amount of feedback used over time when applied to the DBS and the DUC'04 corpus. We can see from the figure that all models show an improvement of +.45 ROUGE-2 after merely 4 iterations

[7] We prune bigrams consisting of only functional words.