프로젝트 자체 성공 평가 척도

17조 summer팀

요약된 문장이 얼마나 요약이 잘 되었는지 즉, 신뢰할 수 있는 요약문인지에 대한 평가를 위해 3가지 방법을 생각해보았습니다.

첫째, 요약되기 전 텍스트에 문맥에 어긋나는 텍스트를 임의로 삽입합니다. 이 문장이 요약된 텍스트에서 나타나지 않는다면 요약이 잘 되었다고 볼 것입니다.

둘째, 사용자에게 별점으로 평가를 받을 것입니다. 평균 별점이 5점만점에 3점 이상일 때 요약이 잘 되었다고 판단할 계획입니다. 별점 3점 이상은 60%이상의 만족도를 보인다는 것이며 과반수 이상의 수치이므로 요약이 잘 되었다고 판단할 수 있습니다.

셋째, 계속 이전의 알고리즘과 수정된 알고리즘을 비교하여 많이 개선이 되었다면 어느 정도 성공을 이루었다고 생각할 것입니다. 알고리즘 간의 비교는 ROUGE 평가 방법을 사용할 계획입니다.

ROUGE (Recall Oriented Understudy of Gisting Evaluation)

ROUGE의 아이디어를 간략하게 정리하자면, 먼저 모범 답안 요약문을 만들어놓고 자동 요약된 요약문이 모범 답안과 얼마나 유사한지를 비교하는 것입니다.

ROUGE-n

The cat was on the my keyboard (인간이 만든 모범 답안)

Cat sat on the keyboard (요약 알고리즘 결과)

ROGUE-1 에서는 1-gram, 즉 단어 한 개의 단위로 두 요약문을 비교합니다.

{ the, cat, was, on, my, keyboard }

{ cat, sat, on, the, keyboard }

Precision = (양쪽 모두 일치하는 단어 개수} / (모범 답안의 단어 개수)

Recall = (양쪽 모두 일치하는 단어 개수) / (자동 요약 결과의 단어 개수)

여기서 Precision = 4 / 6, Recall = 4 / 5 가 됩니다.

저희는 ROUGE-1, ROUGE-2, ROUGE-SU 3가지 방식을 사용해 알고리즘의 정확도를 수치화를 합니다. 모범 답안을 잘 뽑은 뒤에 그거에 비해 얼마나 잘 요약이 이루어졌는지를 판단할 수 있고, 전 알고리즘과의 비교로 정확도가 개선되었다는 걸 보여줄 생각입니다.

A	B	C	D	E	F
ROUGE-Type	Task Name	System Name	Avg_Recal	Avg_Precis	Avg_F-Sco
ROUGE-L+S	TASK1	SYSSUM1.TXT	0.52318	0.67521	0.58955
ROUGE-1+S	TASK1	SYSSUM1.TXT	0.50472	0.69481	0.5847
ROUGE-2+S	TASK1	SYSSUM1.TXT	0.43889	0.59398	0.50479
ROUGE-SU4	TASK1	SYSSUM1.TXT	0.45	0.62718	0.52402