

문서 추출 품질의 계측에 관하여

문서기술연구 제12권 2호 (2026) 발췌

1. 문제

레이아웃 모델은 페이지를 이미지로 보고 본문을 찾는다. 그래서 본문이 아니라고 판단한 영역은 결과에서 통째로 사라진다. 사라진 것은 로그에 남지 않는다 — 없는 줄은 세어지지 않기 때문이다.

원문에서 무손실로 뽑은 텍스트가 있으면 이야기가 달라진다. 그 텍스트를 정답지로 두고 추출 결과와 대조하면, 무엇이 얼마나 빠졌는지 숫자로 말할 수 있다.

2. 커버리지

페이지 p 의 커버리지는 원문 텍스트가 추출 결과에서 발견되는 비율로 정의한다.

$$C(p) = |R(p) \cap E(p)| / |R(p)|$$

여기서 $R(p)$ 는 원문 기준 토큰 집합, $E(p)$ 는 추출 결과의 토큰 집합이다. 엄격한 문자열 유사도 대신 집합 포함 관계를 쓰는 이유는 순서 차이에 별점을 주지 않기 위해서다. 표를 읽는 순서는 모델마다 다르지만, 그것이 손실은 아니다.

페이지 경계를 넘나드는 문단 때문에 대조 창은 ± 3 페이지로 둔다. 창을 좁히면 정상 추출도 손실로 잡힌다.

3. 숫자 전수 대조

커버리지만으로는 부족하다. 금액 한 줄이 빠져도 페이지 전체 커버리지는 0.98을 넘길 수 있다. 문서에서 의미가 가장 무거운 토큰은 숫자이므로, 숫자는 따로 전수 대조한다.

판정	커버리지	숫자 손실	조치
정상	≥ 0.90	없음	없음
경고	≥ 0.90	있음	구간 보강
저품질	< 0.90	무관	페이지 대체

보강은 손실보다 중복을 택한다. 같은 문장이 두 번 나오는 것은 읽는 쪽이 걸러낼 수 있지만, 없는 문장은 되살릴 수 없다.

4. 남은 것

원문 텍스트를 뽑을 수 없는 문서 — 스캔본 — 에는 이 방법을 쓸 수 없다. 이 경우의 계측은 후속 과제로 남긴다.

1) 토큰은 공백을 제거한 뒤 2-gram으로 자른다.