ACORD: Atticus Clause Retrieval Dataset

Atticus Clause Retrieval Dataset (ACORD) is the first retrieval benchmark for contract drafting fully annotated by experts. It includes 114 queries and over 126,662 query-clause pairs, each ranked on a scale from 1 to 5 stars. The task is to find the most relevant precedent clauses to a query. ACORD focuses on complex contract clauses such as Limitation of Liability, Indemnification, Change of Control, and Most Favored Nation.

FORMAT

ACORD consists of query-clause pairs provided in BEIR format. However, unlike in standard BEIR format, irrelevant clauses are explicitly rather than implicitly annotated. This reduces the incidence of false negatives. See the paper appendix for details.

- BEIR Format: In the BEIR format for ad-hoc IR datasets, the clauses and queries can be found in corpus.jsonl and queries.jsonl, respectively. Relevance scores for the train, development, and test splits can be found in qrels/{train,test,valid}.tsv. Relevance scores range between 0 and 4 (corresponding to lawyer ratings from 1 to 5 stars).

- Query-Clause Pairs: Each pair consists of an attorney-written query and a corresponding clause from the corpus. Queries are crafted to represent legal drafting needs, while clauses are annotated with a relevance score (0-4) based on how well they address the query.

Train/Validation/Test Splits

- Training Set: 45% of queries.

- Validation Set: 5% of queries.

- Test Set: 50% of queries.

Each split ensures representation across all clause categories for balanced evaluation.

The dataset includes the following supplemental materials created by legal experts that are provided for interested researchers to conduct additional experiments, however, they were not used in our experiments:

- 2-5 Star Query-Clause Pairs in csv file with more nuanced scores (10 different scores instead of 4 scores); and

- Queries in short, medium and long formats.

DATA STATISTICS

- Total Query-Clause Pairs: over 126,662.

- Clause Categories: 9.

- Limitation of Liability, Indemnification, Affirmative Covenants, Restrictive Covenants, Term, Governing Law, Liquidated Damages, Third-party beneficiary, IP ownership/license.

- Rating Scale:

- 5: Highly relevant clauses addressing all key aspects of the query.

- 4: Relevant clauses addressing most aspects of the query, with minor omissions.

- 3: Partially relevant clauses addressing some aspects of the query but missing key details.

- 2: Non-relevant but helpful clauses within the same category as the query.

- 1: Irrelevant clauses from a different category, unrelated to the query.

All clauses, including irrelevant zero-score clauses, are explicitly are annotated to avoid false negatives, ensuring comprehensive and high-quality data.

TASKS

ACORD is an ad-hoc frames clause retrieval as an ad hoc information retrieval task. Given an attorney-written query, the goal is to rank clauses by relevance.

- Metrics: NDCG@5, NDCG@10, 3-star precision@5, 4-star precision@5, 5-star precision@5.

ANNOTATION PROCESS

The annotation process follows a five-step methodology:

1. Extraction: Relevant clauses are identified by trained annotators.

2. Retrieval: Annotators locate 10 relevant and 20 2-star clauses per query.

3. Scoring: Queries are rated by attorneys using detailed guidelines.

4. Reconciliation: Disputes are resolved by a panel of legal experts.

5. Expansion: Additional irrelevant clauses are added to ensure comprehensive coverage.

Annotations are reviewed by legal professionals to ensure consistency, accuracy, and reliability.

CLAUSE SOURCES

ACORD’s Contract Corpus is derived from:

1. EDGAR contractsSystem: Publicly available filings from the U.S. SEC.

2. Terms of Service contracts: Published by selected Fortune 500 companies.

Annotations include clauses extracted by legal experts, ensuring contextual relevance to the queries.

LICENSE

ACORD is licensed under Creative Commons Attribution 4.0 (CC BY 4.0), allowing both commercial and non-commercial use.

CONTACT

More information can be found at atticusprojectai.org/acord.

For questions, suggestions, or further information, contact the authors at wang.steven.h@gmail.com, [wei@attiscusprojectai.org](mailto:wei@attiscusprojectai.org), or aplesner@ethz.ch.