MERGER AGREEMENT UNDERSTANDING DATASET

Merger Agreement Understanding Dataset (MAUD) v1 is a reading comprehension dataset with over 39,000 examples and over 47,000 annotations based on legal text extracted from 152 publicly available merger agreements involving publicly traded target companies. MAUD answers 92 questions in each merger agreement.

MAUD is curated and maintained by The Atticus Project, Inc. to support NLP research and development in legal contract review. It is created under the supervision of experienced lawyers and used by the 2021 American Bar Association (ABA) Public Target Deal Points Study, or the ABA Study. Code for replicating the results and the trained model can be found at https://github.com/TheAtticusProject/maud.

===

FORMAT

MAUD contains three subsets (main, additional, and counterfactual) corresponding to the three methods by which annotations were generated, each described in details below. We release MAUD as three separate CSV files: MAUD_train, MAUD_dev and MAUD_test, representing the dataset splits used by us to generate the baseline benchmark.

MAUD contains 8,226 unique text annotations and 39,231 question-answer annotations (i.e. examples), for a total of 47,457 annotations.

Each row of the CSVs is an input-output example that can be used to train or evaluate ML models.

The columns of the CSVs are: data_type, contract_name, text, answer, label, question, subquestion, text_type, and category.

* data_type: The method by which the example was generated, either “main”, “additional”, or “counterfactual”.
* contract_name: Either an anonymized string representing the contract from which the deal point text was sourced, or “<COUNTERFACTUAL>”, meaning that the deal point text was created by editing an original deal point text.
* text: The deal point text in this example.
* answer: The deal point answer in this example.
* label: For convenience, an integer label corresponding to the answer value.
* question: The deal point question in this example.
* subquestion: If the deal point question is multiple choice, then ‘<NONE>’. If the deal point question is multilabel, then this is the value of a multilabel answer.
* text_type: A category describing the type of text extraction (different from deal point category).
* category: The deal point category for the question.

- main dataset: This subset contains 20,623 examples with original deal point text extracted from 152 EDGAR merger agreements by expert annotators. The deal point text below is truncated for display. An annotator's task is to pick the correct answer(s) using the deal point text.

data_type	contract_name	text	answer	label	question	subquestion	text_type	category
main	contract_1	"Company Material Adverse Effect" shall mean any state of facts, change, condition, occurrence, effect, event, ...	"Would" (reasonably) be expected to	0	FLS (MAE) Standard-Answer	<NONE>	MAE Definition	Material Adverse Effect
main	contract_2	(i) each share of Company Common Stock (including each share of Company Common Stock described ...	All Cash	0	Type of Consideration	<NONE>	Type of Consideration	General Information
main	contract_3	Section 3.1 Organization, Standing and Power. <omitted>Section 3.2 Capital Stock. <omitted>(b) All outstanding shares of capital stock and other voting securities or ...	Authority	1	Accuracy of Fundamental Target R&Ws-Types of R&Ws	Authority	Accuracy of Target R&W Closing Condition	Conditions to Closing

- abridged dataset: This subset contains 14,928 examples with deal point text extracted from 94 of the 152 merger agreements included in the main dataset. In the additional dataset, deal point texts are abridged to delete portions of legal text in the main dataset that are not pertinent to the deal point question. Because many deal point texts contain answers to multiple deal point questions, we provide the abridged texts to guide a model to recognize the most pertinent text.

- rare answers dataset: This subset contains 3,680 annotations that have rare answers to a question. Legal experts made small edits to texts in the main dataset to create a deal point with a rare answer. This ameliorates the imbalanced answer distribution problem prevalent in the main dataset.

The number of examples by splits (train, dev, test) and by dataset (main, counterfactual, additional) is listed below:

	Train	Dev	Test	Overall
main	13,256	3,471	3,896	20,623
abridged	9,647	2,526	2,755	14,928

rare answers	2,924	756	0	3,680
overall	25,827	6,753	6,651	39,231

The number of examples in each dataset by category is listed below:

Category	Main Dataset	Rare Answers Dataset	Abridged Dataset	All Datasets
Conditions to Closing	3411	298	4052	7761
Deal Protection and Related Provisions	6491	2280	5937	14708
General Information	152	17	173	342
Knowledge	388	23	258	669
Material Adverse Effect	8816	871	3273	12960
Operating and Efforts Covenant	1,216	191	1054	2461
Remedies	149	0	181	330
All Categories	20764	3680	17984	39,231

==

TRAIN, DEV, AND TEST SPLITS

We construct the train-dev-test split as follows. We reserve a random 20% of the combined main and additional datasets as the test split. The remaining main and additional annotations are combined with the counterfactual data, and then split 80%-20% to form the train and dev splits.

To avoid data leakage due to main dataset and additional dataset examples having overlapping text and the same answer, we place main and additional examples from the same contract in the same split. All splitting is stratified by question-answer pairs.

==

DOWNLOAD

Download MAUD at www.atticusprojectai.org/maud.

===
TASKS

MAUD is a multiple-choice reading comprehension task. The model predicts the correct deal point answer from a predefined list of possible answers associated with each question. Several deal point questions we take from the ABA Study are in fact multilabel questions, but for uniformity we cast all multilabel questions as binary multiple-choice questions. See below for an example of the task.

Input	-->	MODEL	-->	Output
Deal Point Question: FLS (MAE) applies to Deal Point Text: "Material Adverse Effect" means, with respect to any Person, any event, change, circumstance, occurrence or effect that (i) has, or would have, a material adverse effect on the business..."				Deal Point Answers: [*] Business and operation of Target [*] Ability to consummate transaction [ ] No

===
DEAL POINTS & CATEGORY

Each merger agreement contains deal points, which are important legal concepts standardized by the ABA that define when and how the parties to a merger agreement would be obligated to complete an acquisition. The 2021 ABA Study includes approximately 130 different deal points, 92 of which are represented in MAUD. Each deal point in MAUD is associated with deal point text (clauses extracted by annotations from the merger agreement), one or more predefined deal point questions, one or more deal point answers out of a predefined list, and a predefined category. A list of these deal points can be found below.

There are seven categories of deal points in MAUD:

1. General Information. This category includes the type of consideration and the deal structure of an acquisition.

2. Conditions to Closing. This category specifies the conditions upon the satisfaction of which a party is obligated to close the acquisition. These conditions include the accuracy of a target

company's representations and warranties, compliance with a target company's covenants, absence of certain litigation, absence of exercise of appraisal or dissenters rights, absence of material adverse effect on the target company, etc.

3. Material Adverse Effect. This category includes a number of questions based on the Material Adverse Effect definition. Material Adverse Effect defines what type of events constitutes a material adverse effect on the target company that would allow the buyer to, among other things, terminate the agreement.

4. Knowledge. This category includes several questions based on the definition of Knowledge. Knowledge defines the standard and scope of knowledge of the individuals making representations on behalf of the target companies.

5. Deal Protection and Related Provisions. This category describes the circumstances where a target company's board is permitted to change its recommendation or terminate the merger agreement in order to fulfill its fiduciary obligations.

6. Operating and Efforts Covenants. This category includes requirements for a party to take or not to take specified actions between the signing of the merger agreement and closing of the acquisition. The types of covenants include obligation to conduct business in the ordinary course of business, and to use reasonable efforts to secure antitrust approval, etc.

7. Remedies. This category describes whether a party has the right to specific performance.

A list of deal point questions and corresponding answers, grouped by deal point category is shown here: https://pastebin.com/W9JwJuFi

==
SOURCE OF DATA

MAUD includes over 39,000 examples and over 47,000 annotations based on legal text extracted from 152 merger agreements from the Electronic Data Gathering, Analysis, and Retrieval system (EDGAR) maintained by the U.S. Securities and Exchange Commission (SEC). Publicly traded companies are required by the SEC rules to file the public target merger agreements with the SEC through EDGAR. Access to EDGAR documents is free and open to the public. The deal point questions and the list of predefined deal point answers to each were created by experienced M&A attorneys and standardized by the ABA.

==
LABELING PROCESS

MAUD is a collective effort of over 10,000 hours by law students, experienced lawyers, and machine learning researchers. Prior to labeling, each law student must attend 70-100 hours of training that included live and recorded lectures by experienced M&A lawyers, follow a 250-page annotation guideline, and pass multiple quizzes.

To create the main dataset and the additional dataset, the law students then conducted manual review and labeling of the merger agreements uploaded in eBrevia, an electronic contract review tool. On a periodic basis, the law students exported the annotations into reports, and sent them to experienced lawyers for quality check. The lawyers reviewed the reports or the labeled contracts in eBrevia, provided comments and addressed student questions. Where needed, reviewing lawyers escalated questions to a panel of 3-5 expert lawyers for discussions and reached consensus. Students or the lawyers made changes in eBrevia accordingly. Each annotation was verified by three additional annotators to ensure consistency and accuracy. The final annotations were exported into the main dataset and the additional dataset.

To create the counterfactual dataset, legal experts copied deal point text from the main dataset and minimally edited it to derive the rare answers. The annotations were then reviewed by an experienced attorney to ensure accuracy.

We then group the main, additional and counterfactual datasets together and split into the train/dev/test CSV files as described in the FORMAT section above.

==

LICENSE

MAUD is licensed under the Creative Commons Attribution 4.0 (CC BY 4.0) license and free to the public for commercial and non-commercial use.

==

PRIVACY POLICY & DISCLAIMERS

We encourage the public to help us improve MAUD by sending us your comments and suggestions to info@atticusprojectai.org. Comments and suggestions will be reviewed by The Atticus Project at its discretion.

The use of MAUD is subject to our privacy policy https://www.atticusprojectai.org/privacy-policy and disclaimer https://www.atticusprojectai.org/disclaimer.

==

CONTACT

Email info@atticusprojectai.org if you have any questions.

==

ACKNOWLEDGEMENTS

Attorney Working Group

Wei Chen, Ravi Mahesh, Rita Anne O'Neill, Michael G. O’Bryan, Jenny Hochenberg, Charlotte May, Ann Beth Stebbins, Gordon Moodie, Andy Nussbaum, Patricia Vella, Mara Goodman, Eugene Kim, Rachel Sholbohm, Daniel Belke, Bradley King, Julie Siegel, Tommi Williams, Briana Bloodgood, June Hu, Nancy Ruben, Victoria Smallwood, Mimi Wu, Polina Demina, Paul Huble, Hallie Shayder Sacchetta, Kirby Smith, Jason Zhang, Jason Fruchter, Chul Hun Lee, Michael Santos, Susie Toumanian, Daisy Beckner, Hanah Kang, Tyler Rosenbaum, Jonas Marson, John Mills, Zach Genett and Rob Townsend

Law Student Contributors

Florian Bachmann, Dimitry Levkin, Warren Clairbarne, Arianne Marcelin-Little, Will Serio, Maddy Cole, David Meyers, Simran Virdi, Kayla Fedler, Catherine Sakurai, Colton Walker, Adam Arbonies, Blake Christie, Justin Fun, Annaliese Martis, Suzanne Truong, Logan Boyle, Marie-Aime Chet, Isabella Gradney, Samantha Mita, Ge You, Paola Zaragoza Cardenales, Thomas Dana, Elizabeth Kaiser, Zhibiao Peng, Shelby Young, Haruto Cheng, Ali Darvish, Jessica Li, Bryant Riveria, Bochen (Jerry) Zhang, Kern Chhikara, Ali Ebshara, Nicholas Lin, Laura Saitta and Qing (Kelly) Zhou

Technical Advisors & Contributors

Steven Wang, Antoine Scardigli, Leonard Tang, Spencer Ball, Anya Chen, Thomas Woodside, Oliver Zhang and Dan Hendrycks