Leveraging Large Language Models (LLMs) for Data Augmentation and Annotation

May Myo Zin
maymyozin@nii.ac.jp
National Institute of Informatics (NII)

1. Problem and Motivation

- Automatic information extraction in the legal domain is essential for enhancing efficiency and accuracy in legal processes and decision-making.
- Named Entity Recognition (NER) is a widely adopted technique for identifying and extracting specific legal entities, facts, or concepts.
- Obtaining labeled datasets that are large and diverse enough to train robust NER models can be a significant challenge.
- Manual data creation and annotation are cost-intensive and time-consuming, requiring human experts.
- Large language models (LLMs) like GPT-3 and GPT-4 address this challenge by generating and annotating legal data in a human-like manner, offering a cost-effective solution for training robust, domain-specific NER models.

2. Methodology

Initial Data Preparation:

Initial Seed Data: we manually create150 legal case descriptions, particularly those involving sale and purchase scenarios.

Desired Entities: *seller, buyer, potential_buyer, contract_name, purchase_date, purchase_product, purchase_price, minor, threatener, victim, duress_date, rescind_date, rescinder.*

Annotation Format: The following annotation format is considered to be quite user-friendly for both creating and reading annotations.

In the contract negotiation, [John Smith](SELLER) disclosed that he had sold his [vacant land](PURCHASE_PRODUCT) to [David Martin](BUYER) through a [Land Purchase Agreement](CONTRACT_NAME). The agreed-upon price was [$200,000](PURCHASE_PRICE), with the transaction set to be completed on [April 2, 2023](PURCHASE_DATE). Notably, [David Martin](MINOR),the buyer, was a minor at the time the agreement was made. Subsequently, on [April 15, 2023](RESCIND_DATE), [David Martin](RESCINDER)initiated the rescission of the contract.

"[ ]" is used to delineate the boundary of the entity.
"( )" is used to indicate the label of the entity.

Table 1: Entity categories in Sale&Purchase case descriptions.

Entity Category	Definition
SELLER	The party or entity that offers goods, services, or property for sale. In a transaction, the seller is the one transferring ownership or providing the specified goods or services.
BUYER	The party or entity that acquires or intends to acquire goods, services, or property through a transaction. The buyer is the one making the purchase and gaining ownership or use of the specified items.
POTENTIAL_BUYER	An individual or entity that is considering or exploring the possibility of making a purchase.
CONTRACT_NAME	The name or title assigned to a legally binding agreement between two or more parties.
PURCHASE_PRODUCT	The specific item, product, or service that is being bought or acquired by the buyer in a transaction.
PURCHASE_PRICE	The agreed-upon monetary value or consideration that the buyer agrees to pay to the seller in exchange for the purchase of a product, service, or property.
PURCHASE_DATE	The date on which a purchase agreement is executed or the date when ownership of the purchased item is transferred from the seller to the buyer.
MINOR	The party who has not reached the age of legal adulthood, typically under the age of 18, when the agreement was made.
THREATENER	The individual or entity that employs coercion, intimidation, or threats to compel another party to undertake a specific action, such as initiating a contract cancellation.
VICTIM	The party or entity that is subjected to duress, threats, or coercion.
DURESS_DATE	The date when the coercive or threatening circumstances arose, prompting one party to feel compelled to cancel the existing agreement against their will.
RESCINDER	The party or entity that initiates the cancellation or rescission of a contract or agreement.
RESCIND_DATE	The date on which a contract or agreement is officially canceled, revoked, or rescinded.

Data Augmentation:

- As the initial set of 150 samples proves inadequate for the development of a robust NER model, we adopted two approaches for augmenting additional training data: ***manual*** and ***GPT-generated***.
- Through ***manual*** efforts, we created 6,300 new annotated samples from the initial set by introducing entity variations, synonym replacement, and paraphrasing.
- In parallel, we generated 6,300 annotated data samples using ***GPT-3*** and an additional 6,300 samples with ***GPT-4***.

- To instruct the GPT model to generate annotated data, the prompt was structured as follows:

Generate 10 different case descriptions (i.e., summaries) of a purchase agreement in a human-written style, each summarizing details such as the seller, buyer, contract_name, purchase_product, purchase_date, purchase_price, minor (i.e., the party who has not reached the age of legal adulthood, typically under the age of 18, when the agreement was made), threatener (i.e., the individual or entity that employs coercion, intimidation, or threats to compel another party to undertake a specific action, such as initiating a contract cancellation.), victim (i.e., the party or entity that is subjected to duress, threats, or coercion), duress_date (i.e., the date when the coercive or threatening circumstances arose, prompting one party to feel compelled to cancel the existing agreement against their will), rescind_date (i.e., the date on which a contract or agreement is officially canceled, revoked, or rescinded), rescinder (i.e., the party or entity that initiates the cancellation or rescission of a contract or agreement), etc. Also, consider entering into an agreement first and then canceling it due to various reasons, including minor cases, threats, or forceful cancellations. Make sure to incorporate real names, dates, accurate financial information, and other relevant details. Please write in a paragraph and use "[]" and "()" to annotate entities.
Learn annotations from the following example cases:
*[*** an example of an annotated case from seed data ***]*
Please avoid replicating the structure and writing style of the provided examples. Instead, create diverse formats and writing styles, such as human-written purchase case summaries, while ensuring accurate annotations. Incorporate names of individuals, companies, or organizations from various countries across Asia and Europe. Display dates in varied formats and styles, encompassing different conventions. Introduce distinct currency symbols and currency types within each summary, ensuring variety across the information provided. Occasionally express purchase prices in words rather than numbers. Utilize various rationales for contract agreement, cancellation, or termination. Moreover, incorporate a range of verbs or synonyms in the generated summaries to enhance diversity.

BERT-based NER:

- Token classification task: the input to the model is a sequence of tokens (words or sub-word units), and the output is a label assigned to each token, indicating whether it belongs to an entity and, if so, what type of entity it is.
- Instead of training the model from scratch, we employ transfer learning.
- We fine-tune a pre-trained $\text{BERT}_{\text{BASE}}$ model for our specific NER task using corresponding annotated data.
- For each experiment, we conduct multiple training sessions with different hyperparameter configurations. The set that performs best on the validation set is then selected.
- Emphasizing the importance of balanced model generalization to address overfitting and underfitting, we incorporate cross-validation.
- Additionally, we implement Early Stopping to prevent overfitting problems.

Zero-shot GPT-based NER:

- The following prompt is designed to instruct the GPT model on how to extract the desired entities from a user input test case:

Extract the following information from the given case summary:
Seller:
Buyer:
.
.
.
Rescind date (i.e., the date on which a contract or agreement is officially canceled, revoked, or rescinded):
Please provide the information exactly as it appears in the provided summary without additional paraphrasing. If a specific entity type is not present in the case summary, indicate it with "N/A." Ensure accuracy and attention to detail in capturing relevant information.
Case Summary:
*[*** user input test case ***]*

3. Experimental Results

Table 2: NER performance on diverse test sets.

Model	Data Augmentation Approach	Samples	Test Set	Precision	Recall	F1
BERT-based NER	Initial Seed Data (Manual)	150	seenTest	0.92	0.92	0.92
			unseenTest1	0.51	0.50	0.50
			unseenTest2	0.45	0.50	0.47
			unseenTest3	0.16	0.24	0.19
BERT-based NER	Manual	6,300	seenTest	0.96	0.96	0.96
			unseenTest1	0.58	0.56	0.57
			unseenTest2	0.45	0.54	0.49
			unseenTest3	0.21	0.27	0.23
BERT-based NER	GPT-3	6,300 (raw)	seenTest	0.96	**0.97**	**0.97**
			unseenTest1	0.67	0.70	0.68
			unseenTest2	0.61	0.64	0.62
			unseenTest3	0.28	0.48	0.35
BERT-based NER	GPT-4	6,300 (raw)	seenTest	0.93	0.80	0.86
			unseenTest1	0.78	0.58	0.66
			unseenTest2	0.57	0.36	0.45
			unseenTest3	0.41	0.30	0.34
BERT-based NER	GPT-4	6,081 (clean)	seenTest	**0.97**	**0.97**	**0.97**
			unseenTest1	**0.90**	**0.90**	**0.90**
			unseenTest2	**0.68**	**0.70**	**0.69**
			unseenTest3	**0.48**	**0.67**	**0.56**
Zero-Shot-gpt3-NER	-	-	seenTest	0.82	0.90	0.86
Zero-Shot-gpt4-NER	-	-	seenTest	0.92	0.90	0.91

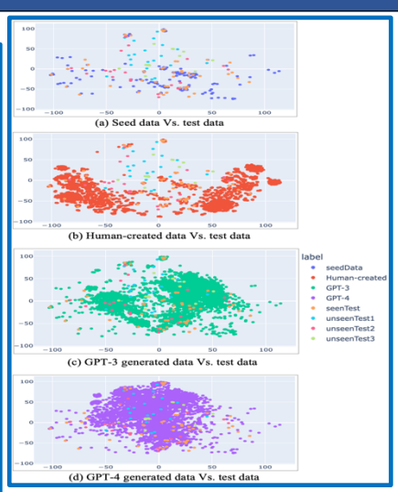

Fig. 1: Augmented training data Vs. test data.

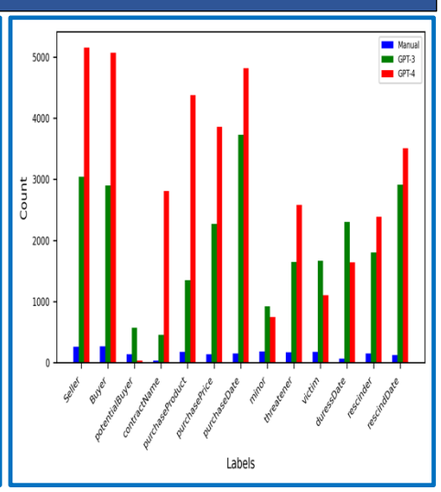

Fig. 2: Comparison of entity counts among three types of augmented datasets.