International Agreements Data Base mining
Project Proposal for NLP Course, Winter 2024

Tomasz Siudalski
Warsaw University of Technology
01161590@pw.edu.pl
Weronika Plichta
Warsaw University of Technology
01194060@pw.edu.pl
Michał Taczała
Warsaw University of Technology
01149437@pw.edu.pl
supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

This project, realized in cooperation with faculty members from the University of Lodz, aims to explore the application of Natural Language Processing (NLP) techniques to analyze a vast collection of international agreements from U.S. states and municipalities. The research will focus on automating the extraction of 13 key attributes proposed by legal researchers, such as areas of cooperation, parties involved, agreement types, and recurring clauses. Leveraging state-of-the-art NLP tools the project will investigate tasks including Named Entity Recognition, relation extraction, or clause frequency analysis and address challenges such as ambiguity and the formal structure of legal documents. By streamlining the analysis process, the project aims to reduce reliance on manual document reviews and also establish a foundation for efficient analysis of similar legal datasets, enhancing the accessibility of actionable insights from complex legal agreements.

1 Introduction

The scientific goal of this project is to analyze a comprehensive database of international agreements concluded by U.S. states to uncover patterns, trends, and key elements of these agreements. These agreements, which cover diverse topics such as

economic development, cultural exchange, and environmental collaboration, provide valuable insights into the evolving dynamics of global partnerships. Traditionally, extracting relevant information from such agreements has been a difficult and time-consuming process requiring expert knowledge. By employing advanced data analysis and NLP techniques, this project seeks to answer several pivotal questions regarding the nature and structure of these agreements. The analysis will focus on identifying key attributes such as areas of cooperation, parties involved, agreement types, and clauses related to duration, extension conditions, and coordination with other entities. Our project also aims to accelerate and streamline the process of information extraction for future agreements enabling faster and more consistent analysis of these documents, eliminating the need for exhaustive manual reviews.

1.1 Research questions

The primary goal of the analysis is to address the following challenges:

- Identification of areas of cooperation mentioned in the agreements.
- Identification of the parties involved (states, institutions, local partners).
- Identification of the types of agreements (e.g., Memorandum of Understanding, Sister Cities Agreement, etc.).
- Determination of the percentage of agreements under the patronage of Sister Cities International.
- Identification of international organizations mentioned in the agreements.
- Determination of the terms of validity for each agreement.
- Identification of the length of each agreement (number of pages or words).
- Determination of the conditions for extending each agreement (automatic/ by decision).
- Analysis of the frequency of recurring clauses in the agreements (always, often, rarely) – the level of detail in the agreements.
- Identification of the partners with whom the agreements tend to be more detailed.
- Indication of whether the agreement includes an evaluation of its implementation.

- Identification of whether the agreement mentions any coordination of activities with other entities (e.g., government, other cities/states, international organizations).
- Identification of whether the agreement refers to other legal documents.

1.2 Significance of the project

This project addresses the need for efficient document analysis in legal and administrative contexts by automating the extraction of critical information from international agreements. By reducing the time and effort required for manual reviews, the methodology enables stakeholders to quickly retrieve actionable insights, identify trends, and make decisions. Furthermore, the methodology of our project can be applied to similar datasets, such as treaties or trade agreements, showcasing the broader applicability and impact of this research.

2 Literature review

Analysis of legal and formal documents like ”paradiplomacy” is quite a different task from the analysis of a standard text. There are a lot of specific words or complex language. However, in the last few years there has been an increase in legal NLP research [4], which causes more data availability and code reproducibility.

To analyze formal text, NLP and computational methods are helpful, to extract information from legal texts. For that fine-tuned models like BERT can handle domain language. Also, Named-entity-recognition and relation extraction can help find relationships between words in documents.[8] What is worth noting is that legal documents often contain ambiguous words that might be hard to analyze. Another potential problem might be that the dataset is unbalanced and contains most of the agreements of a certain type, which might cause the model to learn the most common type of agreement instead of the pattern.

In another study [7] for quite a similar task to paradiplomacy, after annotating data by hired specialists, authors used BERT and Named Entity Recognition for modeling and sentiment analysis. It turned out, that in this type of document, most of the sentences are neutral. A small number of sentences were positive, and few were negative(mostly encountered in the ”unmet goals” part). The preprocessing to achieve such results contained: text extraction (because some PDFs were saved as images), then splitting text into separate paragraphs, spelling corrections(mostly due to unsatisfactory quality of the image-pdf data), and converting to English (because documents were French).

The next study [6] addresses the same problem as the paradiplomacy task. It shows that the framework's use of NER and rule-based extraction is quite useful for handling the structured and specific language that is often used in legal agreements. Since both tasks are very similar, using these techniques to identify agreement types, and international organizations could improve the results of the paradiplomacy project. However, the problem with ambiguity is mentioned in this study once more, as the authors stress this problem and its importance, as it's very common for legal documents and is very likely to appear in paradiplomacy documents as well.

2.1 Blackstone

Blackstone presents a solution for automatically identifying references to legal documents within agreements. Built on spaCy, this open-source legal text processing model is specifically trained to recognize various types of legal references through its Named Entity Recognition (NER) component. It can identify citations, case names, legal instruments (like acts and conventions), specific provisions within those instruments, and court references. While initially trained in UK case law, the model has shown good generalization to other legal systems. However, it's important to note that this is still a prototype with around 70% accuracy (F1 score) for its NER component. Despite these limitations, Blackstone's specialized legal NLP capabilities make it a valuable tool for systematically extracting and analyzing legal document references from agreements, which we want to use in reference to point 13 of our guidelines. Link to GitHub: `https://github.com/ICLRandD/Blackstone/blob/master/README.md`

2.2 LEGAL-BERT

LegalBERT [1] is a family of BERT-based language models specifically trained for legal text processing. Unlike the original BERT, which is pre-trained on general-purpose texts like Wikipedia, LegalBERT adapts to the legal domain by using domain-specific corpora, such as EU and UK legislation, US court cases, and contracts. The training data, spanning 12GB, consists of diverse legal documents that enhance the model's understanding of legal-specific vocabulary and syntax. The authors compared three strategies for BERT in legal tasks: using base BERT, further pre-training BERT on legal corpora, and pre-training BERT from scratch. They showed that the two latter options yield superior performance in legal tasks compared to using the base BERT model by evaluating them on several legal tasks, including multi-label text classification, binary classification, and Named Entity Recognition (NER) for contract elements. LegalBERT is an open-source resource

available on Hugging Face and may be a valuable tool in our research.

2.3 CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review

The Contract Understanding Atticus Dataset (CUAD) [3] is an NLP dataset created to support automated legal contract review as part of The Atticus Project. It contains over 13,000 expert annotations spanning 41 categories of contract clauses, extracted from 510 diverse contracts. These include clauses like governing law, anti-assignment, perpetual licenses, and non-compete agreements. CUAD aims to automate the time-consuming process of extracting key clauses from lengthy contracts which is usually performed manually by legal experts. CUAD serves as a benchmark for assessing NLP models in specialized domains. The authors fine-tuned Transformer-based models such as BERT and DeBERTa [2] on the dataset showing promising but not ideal performance. For example, DeBERTa-xlarge achieves a Precision of 44% at 80% Recall, highlighting substantial room for improvement. The dataset is particularly valuable for its high-quality annotations, which include rigorous quality checks by trained legal professionals. Models fine-tuned on CUAD have the potential to significantly reduce the time and cost of extracting valuable information from contracts.

2.4 Phi-3

Phi-3-mini-4k-Instruct [5] is a 3.8 billion parameter language model developed by Microsoft, designed to balance performance and computational efficiency. Its compact size makes it ideal for projects with resource constraints, such as ours, where scalability and precision are critical. The model employs a data-optimal training approach, utilizing heavily curated web and synthetic data, enabling it to match much larger models like GPT-3.5 on key NLP benchmarks such as MMLU and HellaSwag. Its transformer-based architecture incorporates advanced features like LongRope encoding for extended context handling and block-sparse attention for efficient memory usage. Post-training refinements, including supervised fine-tuning (SFT) and Direct Preference Optimization (DPO), further enhance its capabilities in tasks like reasoning and information extraction. We chose Phi-3-mini-4k-Instruct due to its proven effectiveness in specialized tasks, making it particularly suitable for analyzing complex legal documents within our project's computational constraints.

3 Description of dataset

In a collaborative research initiative with faculty members from the University of Lodz and the University of Warsaw, our team obtained a collection of over 600 legal documents, all sourced from the HeinOnline legal database. These documents represent an array of international agreements, specifically focusing on two main categories: first, formal agreements between individual US states and their counterparts (states or provinces) in other countries, and second, city-to-city agreements such as Memoranda of Understanding and Sister Cities Agreements, which establish cultural and economic partnerships between municipalities worldwide. Initially, each document was preserved in its original form as a scanned PDF or as PDF files containing photographic images that were digitally inserted into the PDF format. To enhance accessibility and enable digital analysis, HeinOnline's database aministrators employed optical character recognition (OCR) technology. This OCR process systematically converts the scanned images into text files.

Variation in the structure and level of detail

A key feature of this document collection is the wide variation in their structure and level of detail. Although some agreements are extensively detailed, containing comprehensive sections on objectives, responsibilities, implementation procedures, and legal frameworks, others are considerably more concise, presenting only basic terms and general principles of cooperation. This heterogeneity in document structure reflects the diverse nature of international agreements, which can range from highly formalized legally binding documents to more informal memoranda of understanding. The variation in detail and structure also appears to correlate with factors such as the scope of cooperation and the jurisdictional level of the participating entities.

Formal language

What's consistent across all these documents, however, is their use of formal, legal language. As official international agreements, they maintain a high level of formality in their writing style, employing specialized legal terminology, complex sentence structures, and standardized diplomatic phrases.

4 Data preprocessing

4.1 PDF to TXT

Data are divided into two main categories: some of them are PDF files with text that can be directly copied and interpreted, and the rest are PDFs created from scans or other image files. We used ORT from tesseract library to retrieve text values from images

4.2 Text Cleaning

The text has been converted to lowercase. Stopwords(like then, moreover, so) have been deleted deleted, and special characters removed.

4.3 Fixing misspelled words

As mentioned in the study [7], Sentences obtained with OCR might contain errors in spelling (especially if the quality of the images is not satisfactory). For the POC stage, we hasn't implemented fixing words yet, but it is important step that will be done in the future if needed.

4.4 Annotating entities

Some entities like countries or state names, had to be annotated manually and in dependence on the task.

5 Exploratory data analysis

Figure 1: Most common words

As we can see on the picture 1, the most common words and phrases in agreements are ”party” and ”united state” which isn’t surprising as we deal with agreements between different American states.

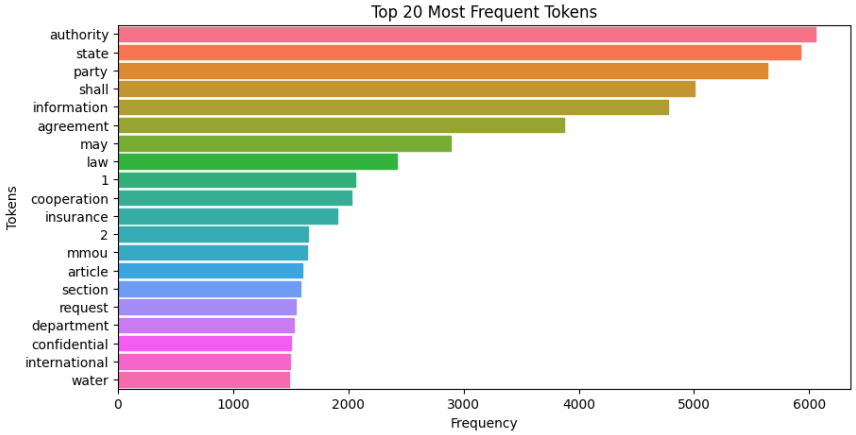

Figure 2: Most frequent words

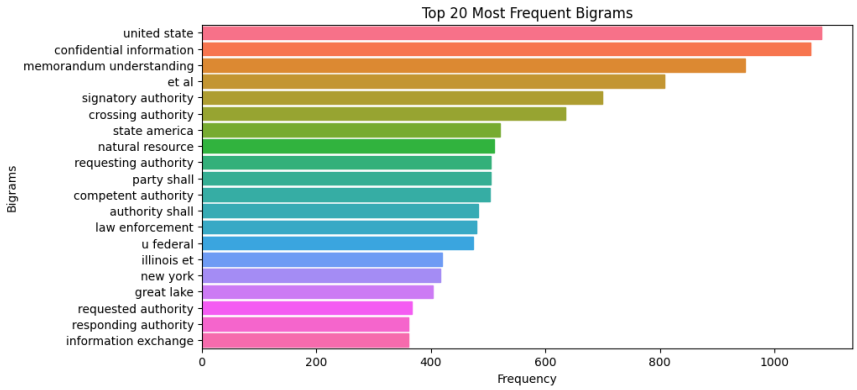

Figure 3: Most frequent bigrams

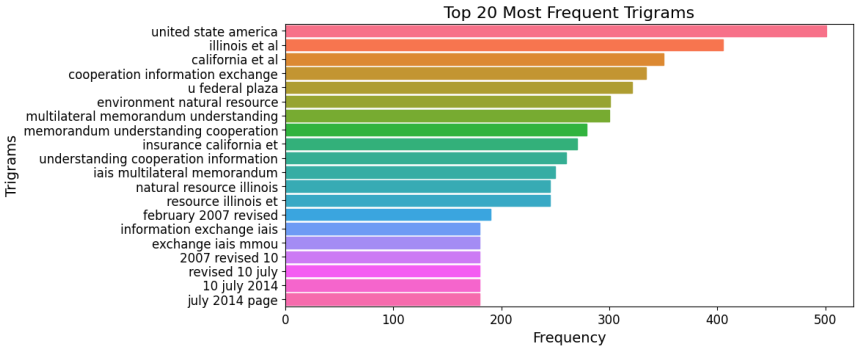

Figure 4: Most frequent trigrams

If we look at the most frequent words, bigrams, and trigrams, we can see some pattern, that a lot of words from the ”most frequent words” plot are present also in the bigram and trigram chart.

Moreover, most of the agreements consist of less than a thousand words 5. There are also some outliers with up to 20 thousand words.

The average length of a word for an article after text-cleaning is 7 letters. Is seems a very big value, but there are 2 reasons for that.

- Most of the stopwords are short (and they have been deleted)
- Agreements are written with formal language which usually consists of longer words

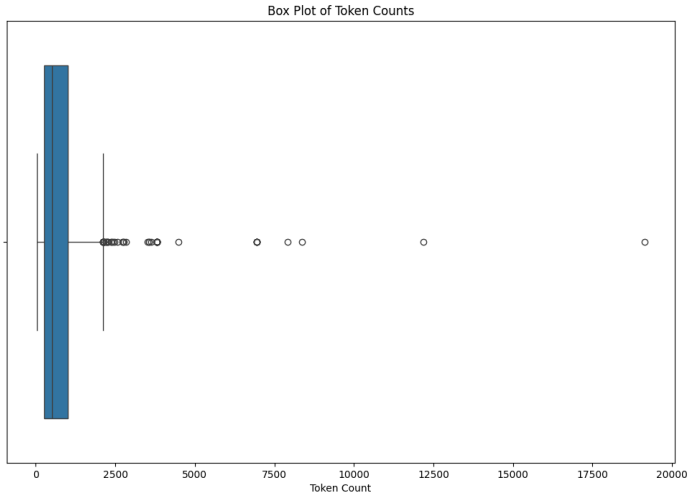

Figure 5: Number of words boxplot

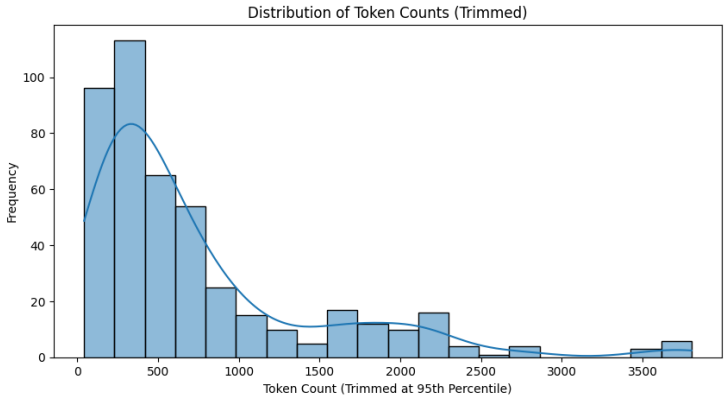

Figure 6: Trimmed number of words

6 POC and experiments

6.1 Identification of areas of cooperation

To identify areas of cooperation within international agreements, we utilized the Phi-"3-mini-4k-Instruct" language model. Due to the model's constraint of a 512-

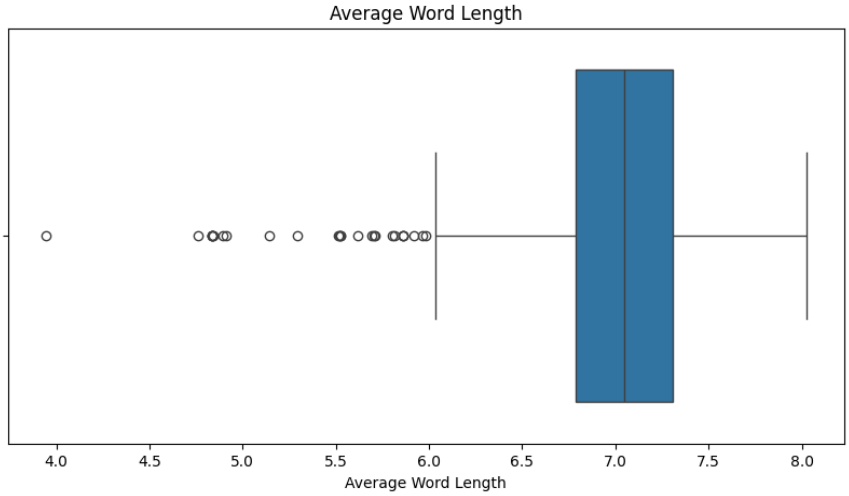

Figure 7: Caption

token context window, processing entire documents at once was not feasible. Segmenting the documents into smaller chunks led to overwriting previously identified areas and significantly increased processing time.

To address these limitations, we adopted a more efficient approach using the KeyBERT model. KeyBERT leverages BERT embeddings to extract keywords quickly, making it well-suited for our task. We generated 20 key phrases from each document, capturing its central themes and concepts. These extracted phrases were subsequently provided as input to the Phi model, enabling it to infer potential areas of cooperation more effectively.

While this method may overlook subtle or minimally mentioned details, it ensures that the primary themes of the documents are identified. The generated cooperation areas appeared reasonable based on preliminary evaluations. However, expert validation would enhance the reliability of the results and support further refinement of the model's reasoning capabilities.

6.2 Identification of the parties involved

We employed the same Phi model for the task of extracting organizational names, encountering similar limitations due to its constrained context window. Despite extensive prompt tuning, the model's outputs were inconsistent, occasionally returning too few or too many organizations and sometimes generating hallucinated results.

To enhance extraction accuracy, we integrated the popular named entity recognition (NER) model SpaCy, which effectively identifies entities labeled with the ”ORG” tag in the text. Although SpaCy's results contained duplicates and substantial noise, feeding these outputs into the Phi model enabled better filtering and identification of legitimate organizations.

The extracted organization names and abbreviations were stored in JSON format. Overall, the extracted organizations were mostly accurate, though some results required manual verification to filter out misidentified entities such as document titles or unrelated terms.

6.3 Identification of types of agreements

For this task, we used a ”facebook/bart-large-mnli” model from the transformer library. This implementation yielded divided agreements into several subgroups for each type of agreement.

6.4 Identification of percentage of Sister Cities patronage

The same as above, we used a ”facebook/bart-large-mnli” model from the transformer library for the zero-shot-classification task. This model yielded 24% of agreements under the Sister Cities International Patronage.

6.5 Identification of international organizations

The task was combined with 6.2.

6.6 Determination of the terms of validity for each agreement

For this task, we also utilized the Phi model, processing documents in chunks due to its limited context window. To identify relevant date-like entities, we employed the SpaCy named entity recognition (NER) model. Detected dates were then passed to the language model along with the chunk of text, prompting it to determine whether each date referred to an agreement's expiration. If an expiration date was confirmed, the process for that document was halted, and the date was extracted.

Since dates appeared in various formats, including textual representations and different numeric formats, implementing a unification step could improve consistency and facilitate further processing.

6.7 Identification of length of each agreement

The task was to count all the words from each agreements. Because during data preprocessing we distilled words, we just counted them without using any external model.

6.8 Determination of the conditions for extending each agreement

We use Phi to identify agreement extension conditions, detecting whether renewals occur automatically or require explicit decisions. The model analyzes texts for specific extension clauses, renewal terms, and decision requirements. The implementation faces similar text chunking limitations - agreements exceeding context window cannot be fully processed.

6.9 Analysis of the frequency of recur- ring clauses in the agreements

Counting words was implemented straight forward with summing functions.

6.10 Identification of the partners with whom the agreements tend to be more detailed, how text is detailed.

As we didn't find any accurate metrics for measuring detail in the formal agreements, we've merged this task with next one, that is responsible for evaluation of implementation for such an agreement. We assume that when the agreement provides a given specific plan for realization of cooperation plan, we can treat it as detailed agreement.

6.11 Indication of whether the agreement includes an evaluation of its implementation.

We use Phi to identify detailed cooperation plans within agreements, specifically focusing on detecting organized activities like meetings, collaborative events, and joint initiatives and evaluating whether these planned objectives were successfully implemented. Current limitations include the absence of text chunking functionality, meaning agreements exceeding the model's context window cannot be fully processed.

6.12 Identification of whether the agreement refers to other legal documents.

The Blackstone Named Entity Recognition (NER) model, pre-trained on 70,000 UK legal documents, was implemented to detect references to external legal documents. This model was selected due to its training on formal legal language similar to that found in paradiplomacy agreements. Specifically we've implemented Blackstone's Named Entity Recognition (NER) capability, specifically utilizing its CITATION entity detection for identifying legal document references. Currently, the system detects references in a specific format containing both the date and legal document code.

References

[1] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school, 2020.

[2] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021.

[3] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review, 2021.

[4] Daniel Martin Katz and et al. Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*, 2023.

[5] Microsoft. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[6] Amin Sleimi et al. An automated framework for the extraction of semantic legal metadata from legal texts. *arXiv preprint arXiv:2001.11245*, 2020.

[7] Joanna Wojciechowska, Mateusz Odrowaz-Sypniewski, Maria W. Smigielska, Igor Kaminski, Emilia Wiśnios, Bartosz Pieliński, and Hanna Schreiber. Deep dive into the language of international relations: Nlp-based analysis of unesco's summary records. In *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*, pages 75–87. Association for Computational Linguistics, 2023.

[8] Anna Wróblewska, Bartosz Pieliński, Karolina Seweryn, Sylwia Sysko-Romańczuk, Karol Saputa, Aleksandra Wichrowska, and Hanna Schreiber.

Automating the analysis of institutional design in international agreements. In *Computational Science – ICCS 2023*, pages 59–73. Springer, 2023.

A Preliminary results

Agreement	Cooperation Areas
1	Energy Management, Clean Energy Innovation, New Energy Development
2	Great Lakes Environmental Protection, Basin Research Collaboration, Lake States Governance Coordination
3	Environmental Protection in Shanghai, Climate Environment Cooperation Shanghai-California, Ecological Collaboration between Shanghai and California
4	Inter-Border Coordination, Bilateral Conference, State Participation, Joint Meetings, Mexico States Collaboration
5	Goodwill Delegation Meetings, Joint Economic Cooperation, State-to-Province Collaboration
6	Technology Innovation, Energy Resources Development, Clean Technology Partnership
7	Cross-border Coordination, Border Infrastructure Development, Port Management and Entry Regulation
8	Mutual Benefit Recognition, Clean Energy Research Collaboration, Environmental Protection Funding Allocation
9	Research Collaboration, Vehicle Regulation, Environmental Protection
10	Annexes and Understandings, Insurance Supervision, Multilateral Memorandums

Table 1: Identified cooperation areas

Entity Name	Abbreviation
The United States	USA
Nevada Organized Crime Commission NOCC	-
Arizona Corporation Commission	ACC
National Research Council of Canada	NRC
Government of the State of Arizona	-
Arizona Department of Transportation	ADOT
Sonora Megaregion	-
Arizona Commerce Authority	-
Smart Borders Initiative	-
Department of Infrastructure and Urban Development of Sonora	-
Arizona-Sonora Commission	-
The Danish Ministry of Energy	-
New York State Department of Public Service NYCRR	NYCRR
Freedom of Information Law, New York State	-
The Working Group on Offshore Wind Energy in the Northeastern United States	-
New York State Energy Research and Development Authority NYSERDA	NYSERDA
Utilities Regulatory Commission of New Hampshire, Inc	-

Table 2: Identified entities and their abbreviations from international agreements.

Agreement	Expiration Date
1	September 11, 2025
2	Three years from the date of signature
3	Valid for a five-year period from the signing date
4	Until terminated by the Parties
5	Valid for three years upon the date of signature
6	December 2025
7	September 14, in two equal periods of two years each
8	A four-year period from the date of its signature
9	October 10, 2027
10	Not specified

Table 3: Identified expiration dates for international agreements.