Can Robots Write Treaties? Using Recurrent Neural Networks to Draft International Investment Agreements

Wolfgang ALSCHNER[a,1] and Dmitriy SKOUGAREVSKIY [b]
[a] *Post-doctoral Researcher in International Law, Graduate Institute of International and Development Studies & World Trade Institute*
[b] *PhD Candidate in International Economics, Graduate Institute of International and Development Studies & European University at St. Petersburg*

Abstract. Negotiating international investment agreements is costly, complex, and prone to power asymmetries. Would it then not make sense to let computers do part of the work? In this contribution, we train a character-level recurrent neural network (RNN) to write international investment agreements. Benefitting from the formulaic nature of treaty language, the RNN generates texts of lawyer-like quality on the article-level, but fails to compose treaties in a legally sensible manner. By embedding RNNs in a user-controlled pipeline we overcome this problem. First, users can specify the treaty content categories *ex ante* on which the RNN is trained. Second, the pipeline allows a filtering of output *ex post* by identifying output that corresponds most closely to a user-selected treaty design benchmark. The result is an improved system that produces meaningful texts with legally sensible composition. We test the pipeline by comparing predicted treaties to actually concluded ones and by verifying that our filter captures latent policy preferences by predicting the outcome of current investment treaty negotiations between China and the United States.

Keywords. Recurrent neural network, investment treaties, machine learning, legal drafting, text-as-data, artificial intelligence.

1. Introduction

Negotiators of international investments agreements (IIAs) have a difficult job. Not only do they need to align new bilateral or regional treaties with past practice to avoid inconsistent interpretations – a daunting task in itself given the universe of over 3000 existent IIAs [1]. But they also need adapt these agreements to an ever-changing legal and political environment. The latter has become particularly important with opposition growing in large parts of the world against new investment agreements such as the Transpacific Partnership hotly contested in the United States or their enforcement through investor-state arbitration controversially debated in Europe. Would it not be

[1] Corresponding Author, Center for Trade and Economic Integration (CTEI), Graduate Institute of International and Development Studies (IHEID), Maison de la Paix, Chemin Eugène-Rigot 2, Geneva, Switzerland, Email: wolfgang.alschner@graduateinstitute.ch. We gratefully acknowledge the funding support from the SNF project "Convergence versus Divergence? Text-as-data and Network Analysis of International Economic Law Treaties and Tribunals", from the SNIS project "Diffusion of International Law: A Textual Analysis of International Investment Agreements", and from NCCR trade regulation.

Electronic copy available at: https://ssrn.com/abstract=2984935

beneficial to farm out part of the task of balancing language and interests in international negotiations to artificial intelligence? Recurrent Neural Networks (RNNs) offer the opportunity to do exactly that. With adequate training data, RNNs can teach computers to write treaties. But how well do they perform that task? In this article we train RNNs to draft investment agreements. Our research suggests that RNNs can produce high quality, lawyer-like output. Yet robots are not to replace negotiators any time soon. Our research also shows that composing balanced treaties is more art than algorithm. The most promising avenue for future applications of RNNs in legal drafting is therefore likely to lie in human-machine interactions, where RNNs facilitate bilateral or multilateral negotiations by distilling a first draft from existing practice guided by human input and conditioned by human-imposed filters. This paper is accompanied by an online appendix at http://mappinginvestmenttreaties.com/specials/rnn-experiment.

2. Literature

Partial automation of contractual drafting has a long history in legal informatics. From legal expert systems that emulate contracting steps to document assembly tools that create contracts based on user-entered information, artificial intelligence is used to partially automate contract production [2]. In contrast, fully automated systems that draft agreements without relying on pre-defined rules or human guidance have yet to emerge fully. In other domains, meanwhile, such fully automated systems have emerged with deep learning algorithms being deployed to write Shakespearian plays or Wikipedia entries entirely without human supervision [3]. These machine-learning tools draft novel text chunks based on a corpus of training data. Unaware of any existing, equivalent applications of such tools in the context of legal drafting, this contribution assesses whether such deep-learning approaches can be harnessed to create a fully automated legal document production pipeline from a corpus of training data.

3. Motivation

Why should robots write treaties in the first place? First of all, as with all processes of automation, greater delegation of tasks from humans to computers promises time-efficiency gains. Already today, treaty negotiators look for ways to make treaty drafting more efficient through boilerplate agreements or copy-and-pasting from one treaty to the other. Fully automated drafting would thus be an additional step in the same direction. Second, task complexity warrants greater automation. In preparation of a multilateral treaty, for instance, negotiators may want to start working on a text that condenses their respective practices into a single document. Since distilling a new agreement out of 3000 existing ones is much easier for a computer than for a human, computer-led drafting can be used for such consolidation. A final motivation is normative in nature: computer-led drafting can alleviate power asymmetries in international negotiations. Our earlier research has shown that powerful states are more successful than poorer ones in aligning negotiation outcomes with their prior practice by basing talks on their model templates [4]. Computer-led drafting offers a potential alternative. Rather than starting bargaining based on a treaty template provided by one of the two sides, an automatically generated treaty text that consolidates elements of

Electronic copy available at: https://ssrn.com/abstract=2984935

both sides can function as baseline for negotiations. Hence, efficiency, complexity and normativity warrant a greater use of machine drafting in investment law.

4. Recurrent Neural Networks

In recent years recurrent neural networks (RNN) have been applied in the context of many natural language processing problems. RNNs differ from feed-forward neural network models by allowing cyclical connections between neurons [5]. This architectural change enables the model to represent sequential information efficiently. However, vanilla RNNs demonstrate inferior performance when it comes to problems with long-term dependencies between units in sequence (due to the vanishing gradient problem). Long Short-Term Memory (LTSM) models [6,7] are a family of RNNs that has been shown to exhibit superior performance in sequence modeling tasks (which include language modeling) [8]. The key differences between LTSM model and vanilla RNNs are that (a) standard neurons are replaced with "memory cell" units capable of storing information for a long period and (b) multiplicative gating units are added to blocks of memory cells to regulate when information is accessible by other blocks or is overwritten. In this study we rely on a character-level LTSM architecture implemented by J. Johnson [9], which, in turn, relies on the model of A. Karpathy [3].[2] The LTSM predicts the most probable next character given the input character. Training is done by looking at the discrepancy between the predicted and actual character in the training set and by updating the model to minimize this difference. With a trained model at hand, one can specify an input string and sequentially predict new characters.

5. Using RNNs to write investment treaties

5.1. Dataset and model specifications

Our analysis is based on an English-language full text dataset of 1628 bilateral investment treaties (BITs) collected as part of our earlier work [4]. Each treaty text is split into its article components. We then concatenate the split article texts back to one large text file, preserving the treaty names and parties in treaty headers that precede each treaty, as well as article numbers and names in headers, which precede each article text within each treaty. In total, this procedure yields a corpus of 27,365,615 characters. We use that corpus to build a 2-layer LSTM with 768 nodes per layer, sequence length of 250 characters and a dropout factor of 0.5 to train it on 80% of the data (10% were used for validation and test sets). After 77,000 iterations we achieved a cross-validation loss of 0.2540 on the validation set, while the train set loss was 0.134. Then we specified the starting sequence of "====" (signifies a new treaty delimiter) and generated 150 strings of 100,000 characters each (15 million symbols in total) from the trained model with a temperature of 0.5 (a factor between 0 and 1 by which the predicted character probabilities are divided to supply more innovative results). We then split the generated strings into 770 BITs on the "====" delimiter and uncovered the associated countries from the header lines that the model learned to create after the delimiter.

[2] In what follows, we refer to LTSM as "RNN".

5.2. Results from full-length model

The results from the trained model were encouraging. We were surprised to see that computer-generated treaty provisions were almost indistinguishable from actually negotiated treaty articles both in style and content. Put differently, the algorithm did a very good job in mimicking the work of negotiators when it came to formulating specific clauses. Where the RNN performed much worse than its human counterparts was in composing entire agreements. Predicted treaties often contained more than one clause on the same subject matter creating unwanted redundancies. The output also suffered from the opposite problem, omitting several core treaty elements found in real agreements. Moreover, while actual agreements contain a range of cross-references between commitments, the algorithm either failed to produce such links or created false ones.

To evaluate our method more formally we compared predicted to actually negotiated BITs. Following a procedure we have developed elsewhere and which we have shown to be a useful means to measure similarity between legal documents [4], we disaggregated the RNN-simulated and actual texts into their respective 5-character-gram components and computed their Jaccard distances. Agreements that are closer in style and content tend to be textually closer to agreements that share similar features but farther away from those that do not. The mean of the vectorized matrix of pair-wise Jaccard distances of real treaties was 0.569 (variance 0.006) whereas for the simulated treaties the mean was 0.525 (variance 0.018). We then decomposed this variance by applying principal components analysis to the Jaccard matrices. For the distance matrix of the real texts the first 3 principal components explained 61.5% of variance, whereas for the simulated texts the top-3 components accounted for 85.2% of the variance. The variance in the RNN-generated texts, albeit being more pronounced, was easily explained by 3 principal components. This supports our qualitative assessment that the RNN generated repetitive results.

Finally, the RNN did a poor job of aligning predicted country names with their respective country practice. As we have shown in prior work [4], individual states differ in their approach to investment policy making. This different signature in treaty making was not captured by the RNN. Correlation between the logarithm of country's 2014 per capita GDP and mean Jaccard distance of all the real treaties she signed was -0.642 (p-value < 0.001), whereas for the simulated treaties the correlation was -0.075 (p-value = 0.552). In short, while at first sight, a human would not be able to tell the difference between an actual and a predicted agreement, a closer look would reveal severe shortcomings. The generated output was thus not of the quality needed to live-up to any of the three motivations identified above.

6. Towards a structured RNN pipeline

To boost the performance of our RNN approach, we embedded it in a larger information pipeline that structures and filters the computer-generated results.

6.1. Input side: Article-level training data

To remedy the problem of an improper composition of agreements, we allow users to specify the content of predicted treaties *ex ante*. We do so by first grouping individual

articles by normative categories using an approach we have developed elsewhere [10]. We then let the user choose which of these categories should be contained in the final treaty. Finally, we train the RNN separately on concatenated texts of articles relating to the same normative category and generate the treaty text from the separately trained models split into article-level components. The approach has significant advantages over the previous pipeline, but also important downsides. On the positive side, it allows users to construct a treaty along pre-defined normative categories. On the downside, our training sets are too small to implement the procedure for all treaty features. While categories that are common to virtually all treaties yield training sets that are large enough to properly train the RNN, this is not the case for infrequent categories, which may only be present in a few dozen or hundreds of treaties. We are thus forced to limit our analysis to core investment provisions.

6.2. Output side: Filter based on Jaccard distances

The second innovation we introduce relates to the filtering of the output. We use Jaccard distances to compare each article generated through the RNN to a given benchmark text. We then filter the output documents by selecting only the one that is closest to our benchmark document. Importantly, the use of a benchmark treaty allows us to steer output in a normative direction defined by the user. We can thereby create articles that correspond most closely to the treaty practice of a single country, i.e. articles where the Jaccard distance to a Chinese treaty is smallest. We can also follow the same procedure to identify compromise treaty design, selecting output that minimizes the Jaccard distance between the treaty practices of two countries, say the United States and China. In the latter case, we can specify the relative weight of each country's contribution to the predicting outcome to factor in bargaining asymmetries.

6.3. Evaluation I: Actual vs predicted BITs

To evaluate the performance of our approach, we again first compare predicted to actual BITs. Taking the UK-China BIT (1986) as benchmark, we generated compromise provisions giving each country's practice equal weight. The mean of per-provision Jaccard distances between the real and optimally selected simulated texts was 0.428 whereas the mean of distance between the real and all the simulated provision texts was 0.621. Our structured pipeline thus improved our output. More results and texts are reported at http://mappinginvestmenttreaties.com/specials/rnn-experiment/.

6.4. Evaluation II: Writing the United States-China BIT

We next evaluate the effect of our weighting. For many years, China and the United States have been trying to negotiate a bilateral investment treaty. One of the main areas of disagreement is the scope of investment protection. While the United States has historically been a champion of investor rights, China has been more concerned with preserving host state regulatory powers. A predicted BIT between the U.S. and China thus provides a test case for whether our structured RNN pipeline can capture diverging design preferences. A predicted text weighted in favor of U.S. practice should be more investor-friendly whereas a text weighted in favor of China should be more state-friendly. We selected the closest RNN-predicted articles in terms of Jaccard distance using 100% U.S. and 100% Chinese treaty practice weight as opposing benchmarks.

Encouragingly, the results reflect a progression from more to less investment protection. While the American-centric treaty yields, for instance, a predicted expropriation clause that references indirect expropriation, the Chinese centric treaty only covers directly expropriatory measures. Similarly, the U.S. dominant treaty includes interest payments on top of market value compensation whereas the China dominant agreement only accounts for the value of the expropriated investment. While further testing is needed, this preliminary assessment suggests that our use of *ex post* benchmarks for filtering successfully captures varying treaty design preferences.

7. Conclusion and future work

Can robots write treaties? The answer is a partial yes. We were successful in creating intelligible and legally meaningful computer-generated texts through RNNs. Furthermore, our structured pipeline addressed some of the shortcomings encountered en route. *Ex ante* training on normative sub-categories and *ex post* selection using Jaccard distances to benchmark texts produces not only superior results, but also allows users to guide the process normatively. RNNs may thus have a future in facilitating negotiations, resolving complexities and alleviating power asymmetries.

At the same time, challenges remain. Predicted documents typically lack the internal coherence of human drafted treaties. Furthermore, large training sets are needed to do meaningful prediction. Finally, future evaluations will have to further assess strengths and weaknesses of the pipeline. While the RNNs may perform relatively well on "main stream" language it may fail to provide creative or novel solutions. Yet, even if robots are not going to replace negotiators any time soon, this exploratory research suggests that there is merit in pursuing an RNN-based drafting pipeline. Predicted texts can consolidate past practices into compromise language and thereby serve as a useful starting point for negotiations alleviating power asymmetry concerns. At the same time, predicted treaty texts can enable countries to make contingencies for different negotiation scenarios, prepare multilateral draft agreements or serve as benchmarks that researchers can compare with actual negotiations. RNN-based treaty text prediction thus promises a range of possible future applications.

References

[1] UNCTAD, *World Investment Report 2015: Reforming International Investment Governance*, United Nations, Geneva (2015).

[2] J. Jenkins, What Can Information Technology Do For Law, *Harv. JL & Tech.* **21**(2) (2007).

[3] A. Karpathy. char-rnn. https://github.com/karpathy/char-rnn (2015).

[4] W. Alschner & D. Skougarevskiy, Mapping the Universe of International Investment Agreements, *Journal of International Economic Law* **19**(3) (2016).

[5] A. Graves, *Neural Networks. Supervised Sequence Labelling with Recurrent Neural Networks*, Springer Berlin Heidelberg (2012).

[6] S. Hochreiter & J. Schmidhuber, Long short-term memory, *Neural computation* **9**(8) (2016).

[7] A. Graves & J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, *Neural Networks* **18**(5) (2005).

[8] K. Greff, R. Kumar Srivastava, J. Koutník, B. R. Steunebrink, & J. Schmidhuber, LSTM: A Search Space Odyssey, mimeo, arXiv:1503.04069 (2015).

[9] J. Johnson. *torch-rnn*. https://github.com/jcjohnson/torch-rnn (2016).

[10] W. Alschner & D. Skougarevskiy, Convergence and Divergence in the Investment Treaty Universe – Scoping the Potential for Multilateral Consolidation, *Trade, Law & Development* **8**(2) (2016).