IBM Scenario Planning Advisor: A Neuro-Symbolic ERM Solution

Mark Feblowitz, Oktie Hassanzadeh, Michael Katz,
Shirin Sohrabi, Kavitha Srinivas, Octavian Udrea
IBM Research
{mfeb, hassanzadeh, ssohrab, udrea}@us.ibm.com, {michael.katz1, kavitha.srinivas}@ibm.com

Abstract

Scenario Planning is a commonly used Enterprise Risk Management (ERM) technique to help decision makers with long-term plans by considering multiple alternative futures. It is typically a manual, highly labor intensive process involving dozens of experts and hundreds to thousands of person-hours.

We previously introduced a Scenario Planning Advisor prototype (Sohrabi et al. 2018a,b) that focuses on generating scenarios quickly based on expert-developed models. We present the evolution of that prototype into a full-scale, cloud-deployed ERM solution that: (i) can automatically (through NLP) create models from authoritative documents such as books, reports and articles, such that what typically took hundreds to thousands of person-hours can now be achieved in minutes to hours; (ii) can gather news and other feeds relevant to forces in the risk models and group them into storylines without any other user input; (iii) can generate scenarios at scale, starting with dozens of forces of interest from models with thousands of forces in seconds; (iv) provides interactive visualizations of scenario and force model graphs, including a full model editor in the browser.

The SPA solution is deployed under a non-commercial use license at *https://spa-service.draco.res.ibm.com* and includes a user guide to help new users get started. A video demonstration is available at *https://www.youtube.com/watch?v=Gd4CMKclkBY*.

Introduction

Scenario planning for Enterprise Risk Management, although widely recognized as a critical tool to improve business outcomes (Stulz 1996), has not been widely implemented, especially in small and medium enterprises due to barriers related to the availability of experts and the time and effort involved in creating risk models and scenarios from them (Avanesov 2009; Cardoso and Emes 2014). Typically, the process involves identifying *risk forces* from relevant literature, putting together a *forces causal model* that describes the cause – consequence relationships between risk forces, identifying relevant forces for the company or domain of interest and then describing possible evolutions into the future based on the *forces causal model*, the relevant forces and the interpretation of likelihood, impact, relevance and other dimensions that we want to assign to our forces. When manually executed, scenario planning can only focus on small models (at most tens of forces and relations) and results in few and relatively compact scenarios.

Scenario Planning Advisor

In 2016-2017, in collaboration with the IBM Chief Risk Officer's program for managing emerging risk, we deployed a research prototype internally focused on creating scenarios at scale, assuming risk forces models were already provided by experts in the form of mindmaps (Sohrabi et al. 2018a,b). While this greatly reduced the effort involved in generating scenarios, the effort in creating models remained, estimated by risk practitioners to be in the thousands of person-hours for the average model. Furthermore, our awareness and scenario generation functions still required significant user input or sometimes could not scale for large user requests.

Based on the lessons learned in this deployment, we completely re-designed the Scenario Planning Advisor, with a large focus on usability. Our primary goal was to be able to get users started generating scenarios from a model in minutes to a few hours (compare that to thousands of hours for the manual approach) and have a large percentage (90%) of the generated information considered acceptable or correct by a practitioner. Our secondary goal was to provide a more streamlined and interactive experience, allowing users to interact and customize scenario visualization and automatically extracted forces causal models. We intend to showcase several key features supporting these two key goals.

Model creation. We have largely automated the creation of the forces causal model, previously manually developed by experts as a set of mindmaps. The knowledge engineering process described here is the result of several years of observation and consultation with risk practitioners and is designed to accommodate a range of usage patterns, from quick-start (getting a quick model created within a few minutes to generate scenarios) to detailed (carefully go over model customization). It consists of the following steps:

- The user (practitioner) provides a set seed of risk forces and a set of authoritative sources (PDF or text) such as articles, reports, book chapters, etc., about the domain of interest.
- We created a method to automatically extract a proposed

(subject to human review) forces causal model from the provided documents. The Causal Extraction from Authoritative Sources (*CEFAS*) approach casts the problem of causal relationship detection as a question answering problem using the seed set of forces and the document corpus by asking questions such as *What does X cause?* and *What causes X?*. CEFAS employs Huggingface's Transformers (Wolf et al. 2019) using an ALBERT version 2 xxlarge model architecture[1] (Lan et al. 2020), fine-tuned with SQuAD 2.0 (Rajpurkar, Jia, and Liang 2018). This approach has several benefits on the state of the art, including (i) no training necessary, (ii) detection of cross-sentence causation and (iii) discovery of new candidate forces (open-ended answers).

- After causal extractions completes, an initial draft model is available for customization for the user in three optional steps: (i) forces can be reviewed, renamed or disabled; (ii) equivalent phrasings of forces (e.g., *increase in inflation* and *inflation rises*) are already recommended by *CEFAS*, but can be changed or customized by the user; (iii) a crowd-sourced questionnaire can be run with all users of the model to obtain metadata about the extracted causal relations, such as the impact, likelihood and duration of each; questionnaires are dynamically generated and focus on causal relations the system is unsure about.
- Finally, the user has to select a set of forces to serve as *Implications*, effects to which practitioners pay special attention. We automatically suggest implications to the end user, so this step can range from one click that automatically accepts all the suggestions to a careful, manual curation of this list from the available forces.

Scenario generation. SPA generates scenarios by transforming the forces causal model and a set of forces of interest to the user into a planning domain and problem, applying a planner to obtain multiple (hundreds to thousands) of possible plans, and then clustering these plans – essentially trajectories into the future – into scenarios. We have significantly scaled up our scenario planning capability by switching from a domain specific to a domain independent top-k planner (Katz et al. 2018)[2] and switching to a simpler hierarchical clustering algorithm with a soft time limit. This allows us to generate scenarios in a few seconds, even when starting with very large models (thousands of nodes, tens of thousands of causal relations), and a large number of starting forces (dozens).

Awareness. Our awareness capability automatically gathers relevant articles and feeds related to the risk forces of interests. While previously this required substantial user involvement in topic and keyword modeling, it currently requires no additional user time. Articles are automatically fetched multiple times each day from the *Watson Discovery* service news collection, then semantically re-ranked based on embeddings using BERT-NLI-STSB (Bowman et al. 2015) using the sentence-transformers library and the Google Universal Sentence Encoder (Cer et al. 2018). Together, these models provide us with a small number of relevant articles for each force, which we then cluster into storylines using their embeddings and the HDBSCAN algorithm (Campello, Moulavi, and Sander 2013).

Presentation. We re-implemented our user interface to be a Single Page Application using Angular 9. This allows us to operate easily on a variety of devices, including portable ones. We implemented a number of new Angular components, including a graph rendering component based on *dagre-d3*[3] that can display very large models quickly in the browser. In particular, we use this component to allow users to interact with and manipulate scenario graphical displays, and also implemented a fully-featured model editor to either create models in the browser from scratch or edit automatically extracted ones.

Deployment. We made deployment much easier – most of the SPA solution now deploys as a containerized Kubernetes or OpenShift application, with some NLP services requiring GPU-enabled machines. This allows our small research team to maintain multiple separate deployments easily - one operating internally for over 70 IBM finance teams around the world, an open non-commercial service at *http://spa-service.draco.res.ibm.com*, as well as other licensed services for clients.

Evaluation and Feedback

Between June and September 2020, our internal and external deployments of SPA helped produce more models and scenarios than what risk practitioners would have manually been able to do in years. Currently, these instances are hosting 75 new models created through a combination of automated model creation and manual model editing, with a running average between 4 and 6 scenarios per day. We can contrast this with days to weeks for a team of people to manually create scenarios, even when the causal model is already available.

An external client with a large risk practice has also validated this approach. By running a careful user study using 1240 public company 10k reports, they report that within 4 hours, SPA can process 293 documents producing 1775 cause – consequence relationships, where experts can process 22 documents generating 34 such pairs. Furthermore, they estimate that SPA can generate a model and 10 scenarios in about 11 hours compared to about 3800 person-hours for people to achieve the same. The advantage becomes even more pronounced with each scenario after that, which takes SPA seconds to generate, and hours to days for a person to match.

To conclude, we demonstrate the novel application of neural (NLP) and symbolic (planning) techniques to make scenario planning tractable and accessible to a much wider audience than ever possible before. The approach has been validated with both internal and external commercial deployments, scaling scenario planning capabilities by several orders of magnitude.

[1]Made available by IBM Research for pytorch at https://huggingface.co/mfeb/albert-xxlarge-v2-squad2

[2]This allows us freedom in terms of choosing a different planning flavor, such as diverse planning.

[3]https://github.com/dagrejs/dagre-d3

References

Avanesov, E. 2009. Risk management in ISO 9000 series standards. In *International Conference on Risk Assessment and Management*, volume 24, 25.

Bowman, S. R.; Angeli, G.; Potts, C.; and Manning, C. D. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 632–642. Lisbon, Portugal: Association for Computational Linguistics.

Campello, R. J. G. B.; Moulavi, D.; and Sander, J. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In Pei, J.; Tseng, V. S.; Cao, L.; Motoda, H.; and Xu, G., eds., *PAKDD (2)*, volume 7819 of *Lecture Notes in Computer Science*, 160–172. Springer. ISBN 978-3-642-37456-2.

Cardoso, J. F.; and Emes, M. R. 2014. The Use and Value of Scenario Planning. *Modern Management Science and Engineering* 2(1).

Cer, D.; Yang, Y.; Kong, S.-y.; Hua, N.; Limtiaco, N.; St. John, R.; Constant, N.; Guajardo-Cespedes, M.; Yuan, S.; Tar, C.; Strope, B.; and Kurzweil, R. 2018. Universal Sentence Encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 169–174. Brussels, Belgium: Association for Computational Linguistics.

Katz, M.; Sohrabi, S.; Udrea, O.; and Winterer, D. 2018. A Novel Iterative Approach to Top-k Planning. In *Proceedings of the Twenty-Eighth International Conference on Automated Planning and Scheduling (ICAPS 2018)*. AAAI Press.

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2020. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. In *8th International Conference on Learning Representations, ICLR 2020*, 1–17. OpenReview.net.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* .

Sohrabi, S.; Katz, M.; Hassanzadeh, O.; Udrea, O.; and Feblowitz, M. D. 2018a. IBM Scenario Planning Advisor: Plan Recognition as AI Planning in Practice. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, 5865–5867.

Sohrabi, S.; Riabov, A. V.; Katz, M.; and Udrea, O. 2018b. An AI Planning Solution to Scenario Generation for Enterprise Risk Management. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI 2018)*, 160–167. AAAI Press.

Stulz, R. M. 1996. Rethinking risk management. *Journal of applied corporate finance* 9(3): 8–25.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Scao, T. L.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. M. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. In *Arxiv.org*, 1910.03771.