a large user experience component as to what to do with these identified passages that is still not solved.

9 CONCLUSION

The due diligence problem, identifying different types of passages in documents and quantifying risk associated with them, is the basis of how companies conduct mergers and acquisition. Failures to do this task well can result in dramatic monetary loss. One need only look at HP's $8B loss after acquiring Autonomy for $10B for an example of what can happen if done improperly. In this paper, we have presented and formalized the due diligence problem as an IR task and set it apart from other legal retrieval tasks.

As part of this work, we describe the release of a subset of our internal training data to help foster and encourage active investigation into the due diligence problem. This dataset comprises approximately 4,200 agreements, totaling over 15M sentences, from the US, UK, and Canada annotated for 50 different information needs. Using this dataset, one can not only investigate new methods for conducting due diligence and related problems but can verify and replicate the experiments we have presented herein.

In addition to this dataset, we present our current in-production solution to the due diligence problem, whereby we treat documents as sequences of sentences and use Conditional Random Fields to predict the necessary sentence-level labels for a particular topic. We show that this approach is significantly and substantially better than using a linear classifier and that it achieves substantively better recall when compared to hybrid approaches combining Hidden Markov Models and SVMs. Furthermore, CRFs exhibit less degenerate labelling behaviour than any of the tested approaches.

ACKNOWLEDGMENTS

The authors would like to thank Michael Berner for providing valuable feedback and assistance in editing this work. The authors would also like to thank the numerous annotators who have contributed to making this dataset possible, including, but not limited to, Sondra Rebenchuk and Bettina de Catalogne. The authors would also like to thank the anonymous referees for their valuable comments and helpful suggestions.

REFERENCES

[1] 2017. Legal AI Co.s Seal, Kira + Leverton Show Buoyant Growth. https://www.artificiallawyer.com/2017/09/15/legal-ai-co-s-seal-kira-leverton-show-buoyant-growth/. (Sept. 2017).
[2] Yasemin Altun, Ioannis Tsochantaridis, and Thomas Hofmann. 2003. Hidden markov support vector machines. In *Proc. ICML 2003.*
[3] Simon Attfield and Ann Blandford. 2010. Discovery-led refinement in e-discovery investigations: sensemaking, cognitive ergonomics and system design. *Artif. Intell. Law* 18, 4 (2010).
[4] Jason R. Baron, David D. Lewis, and Douglas W. Oard. 2006. TREC 2006 Legal Track Overview. In *Proc. TREC 2006.*
[5] Jack T. Ciesielski. 2016. How Autonomy Fooled Hewlett-Packard. http://fortune.com/2016/12/14/hewlett-packard-autonomy/. (Dec. 2016).
[6] Charles L.A. Clarke, Gordon V. Cormack, and Elizabeth A. Tudhope. 2000. Relevance ranking for one to three term queries. *Info. Proc. & Man.* 36, 2 (2000).
[7] Cyril W. Cleverdon. 1970. The Effect of Variations in Relevance Assessments in Comparative Experimental Tests of Index Languages. *Cranfield University Technical Report* (Oct. 1970).
[8] Gordon V. Cormack and Maura R. Grossman. 2014. Evaluation of machine-learning protocols for technology-assisted review in electronic discovery. In *SIGIR 2014.*
[9] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *J. Machine Learning Research* 7, Mar (2006).
[10] Susan Dumais. 2016. Keynote at TREC 25th Anniversary. In *Proc. TREC-2015.*
[11] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proc. ACL 2005.*
[12] Bruce Hedin, Stephen Tomlinson, Jason R. Baron, and Douglas W. Oard. 2009. Overview of the TREC 2009 Legal Track. In *Proc. TREC 2009.*
[13] T. Kiss and J. Strunk. 2006. Unsupervised multilingual sentence boundary detection. *Computational Linguistics* 32, 4 (2006), 485–525.
[14] Ben Klaber. 2013. Artificial Intelligence and Transactional Law: Automated M&A Due Diligence. In *ICAIL DESI V Workshop.*
[15] John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. ICML 2001.*
[16] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. *CoRR* abs/1603.01360 (2016).
[17] J. Langford, L. Li, and A. Strehl. 2007. Vowpal Wabbit Open Source Project. Technical Report, Yahoo!. (2007).
[18] Jimmy Lin, Adam Roegiest, Luchen Tan, Richard McCreadie, Ellen Voorhees, and Fernando Diaz. 2016. Overview of the TREC 2016 Real-Time Summarization Track. In *Proc. TREC 2016.*
[19] Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *CoRR* abs/1603.01354 (2016).
[20] Jeffrey Manns and Robert Anderson. 2017. Engineering Greater Efficiency in Mergers and Acquisitions. 72 (Sept. 2017).
[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS 2013.*
[22] Jorge Nocedal. 1980. Updating quasi-Newton matrices with limited storage. *Math. Comp.* 35, 151 (1980).
[23] Douglas W. Oard, Jason R. Baron, Bruce Hedin, David D. Lewis, and Stephen Tomlinson. 2010. Evaluation of information retrieval for E-discovery. *Artif. Intell. Law* 18, 4 (2010).
[24] Douglas W. Oard, Bruce Hedin, Stephen Tomlinson, and Jason R. Baron. 2008. Overview of the TREC 2008 Legal Track. In *Proceedings of the Seventeenth Text REtrieval Conference (TREC 2008).*
[25] Supreme Court of the United States of America. 2017. *Federal Rules of Civil Procedure.*
[26] Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields (CRFs). (2007). http://www.chokkan.org/software/crfsuite/
[27] Nils Reimers and Iryna Gurevych. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proc. EMNLP 2017.*
[28] Adam Roegiest, Gordon V. Cormack, Charles L.A. Clarke, and Maura R. Grossman. 2015. Impact of Surrogate Assessments on High-Recall Retrieval. In *Proc. SIGIR 2015.*
[29] Adam Roegiest and Winter Wei. 2018. Redesigning a Document Viewer for Legal Documents. In *Proc. CHIIR '18.*
[30] James A Sherer, Taylor M Hoffman, and Eugenio E Ortiz. 2015. Merger and Acquisition Due Diligence: A Proposed Framework to Incorporate Data Privacy, Information Security, E-Discovery, and Information Governance into Due Diligence Practices. *Rich. JL & Tech.* 21 (2015).
[31] James A Sherer, Taylor M Hoffman, Kevin M Wallace, Eugenio E Ortiz, and Trevor J Satnick. 2016. Merger and Acquisition Due Diligence Part II-The Devil in the Details. *Rich. JL & Tech.* 22 (2016).
[32] Debbie Stephenson. 2013. Top 10 Due Diligence Disasters. https://www.firmex.com/thedealroom/top-10-due-diligence-disasters/. (Mar. 2013).
[33] Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. 2003. Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering *(SIGIR '03).*
[34] Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. 2004. Support Vector Machine Learning for Interdependent and Structured Output Spaces *(ICML 2004).*
[35] Jyothi K Vinjumur and Douglas W Oard. 2015. Finding the privileged few: Supporting privilege review for e-discovery. *Proc. Ass. Info. Sci. and Tech.* 52, 1 (2015).
[36] Jeroen B. Vuurens and Arjen P. Vries. 2014. Distance Matters! Cumulative Proximity Expansions for Ranking Documents. *Inf. Retr.* 17, 4 (2014).
[37] Robert H. Warren and Alexander K. Hudek. 2017. System and method for identifying passages in electronic documents. (9 May 2017).
[38] William Webber. 2011. Re-examining the effectiveness of manual review. In *Proc. SIGIR Information Retrieval for E-Discovery Workshop.*
[39] Jiashu Zhao and Jimmy Xiangji Huang. 2014. An Enhanced Context-sensitive Proximity Model for Probabilistic Information Retrieval. In *Proc. SIGIR '14.*