Variations in Assessor Agreement in Due Diligence

Adam Roegiest and Anne McNulty
Kira Systems
Toronto, Canada
{adam.roegiest,anne.mcnulty}@kirasystems.com

ABSTRACT

In legal due diligence, lawyers identify a variety of topic instances in a company's contracts that may pose risk during a transaction. In this paper, we present a study of 9 lawyers conducting a simulated review of 50 contracts for five topics. We find that lawyers agree on the general location of relevant material at a higher rate than in other assessor agreement studies, but they do not entirely agree on the extent of the relevant material. Additionally, we do not find strong differences between lawyers who have differing levels of due diligence expertise.

If we train machine learning models to identify these topics based on each user's judgments, the resulting models exhibit similar levels of agreement between each other as to the lawyers that trained them. This indicates that these models are learning the types of behaviour exhibited by their trainers, even if they are doing so imperfectly. Accordingly, we argue that additional work is necessary to improve the assessment process to ensure that all parties agree on identified material.

CCS CONCEPTS

• **Information systems** → **Information extraction**; • **Human-centered computing** → *User studies*;

KEYWORDS

Due diligence, user study, legal retrieval, assessor agreement

ACM Reference Format:
Adam Roegiest and Anne McNulty. 2019. Variations in Assessor Agreement in Due Diligence. In *2019 Conference on Human Information Interaction and Retrieval (CHIIR '19), March 10–14, 2019, Glasgow, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3295750.3298945

1 INTRODUCTION

Assessor agreement in Information Retrieval has traditionally been quite low but its effects on system evaluation have been relatively minimal [2, 17, 18]. In legal electronic discovery ("eDiscovery"), agreement has been higher and variation has attributed more to assessor rather than topical variance [6, 11]. In contrast, due diligence [14], where lawyers review a company's contracts during a merger or acquisition and find the potentially problematic passages, assessor disagreement could yield a potentially over-inclusive report, which would waste senior lawyer time, or miss crucial information, which could lead to a bad deal. Accordingly, understanding how assessor behaviour differs in the due diligence context from other IR contexts allows us to determine the extent to which we can rely on previous results for solutions in due diligence.

This work presents a study we conducted consisting of 9 of our in-house legal professionals, of which four have a high degree of due diligence experience and the remainder have a lower level of experience. The participants were required to find and highlight the five common due diligence topics in 50 short contracts (c.f. Section 3). While this does not necessarily match the scale of actual due diligence projects, it should provide a reasonable barometer of agreement, especially when given the high-costs associated with document review. Furthermore, these participants form what we might think of as an "optimistic" estimate of agreement as they are all familiar with our internal definitions of these topics rather than each reflecting a different institution's conception.

Using these participant annotations, we examine the coarse-grained agreement measures (e.g., Cohen's κ [4] and relevant overlap [18]) that rely on simple overlap of annotations to determine relevance. Such coarse-grained measures may omit valuable information as one participant may highlight more information than another. This type of disagreement (i.e., amount of content), if present, potentially exemplifies differences in assessing strategy that could have non-trivial ramifications (e.g., longer highlights may obfuscate important details). Accordingly, we compare and contrast these coarse-grained agreement measures with measures, influenced by plagiarism detection [12], that account for the amount of (non-)overlapping material between annotations from different participants.

In real-world due diligence, lawyers are often required to review thousands of contracts in very little time which is often not practical. Accordingly, the use of machine learning to train models to identify relevant material is becoming increasingly desirable. Given that premise, we might wonder how users and machine learning models would agree. Following Roegiest et al. [14], we train models for each participant's assessments on a topic. As it is prohibitively too expensive to get participants to review new documents, we use the models as imperfect proxies and have them annotate 20 additional documents. After comparing the agreement of the models on these new documents, we conclude with some thoughts on how to improve model agreement.

CHIIR '19, March 10–14, 2019, Glasgow, United Kingdom

ACM ISBN 978-1-4503-6025-8/19/03...$15.00
https://doi.org/10.1145/3295750.3298945

2 RELATED WORKS

Voorhees' work [18] into assessor agreement at TREC has led to many subsequent investigations about task specific agreement. Despite the fact that the TREC assessors came from similar backgrounds, Voorhees found that relevant assessments overlapped between 0.42 to 0.49 of the time on the TREC 4 collection. In spite of the high disagreement, Voorhees found that system evaluation and rankings were generally robust to changes in assessors. This result has been replicated by different authors [2, 17]. Assessor agreement can be affected by many different factors [1, 8, 19, 22, 24], of which Kinney et al.'s [8] and Al-Harbi and Smucker's [1] finding that assessing instructions may have some affect on behaviour is most applicable to this work.

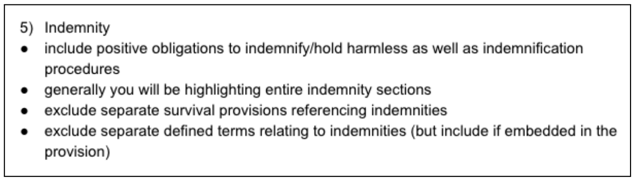
In legal contexts, assessor agreement becomes important when we consider the fact that there can be repercussions when disagreement occurs (e.g., not returning material because one lawyer believes it not to be relevant may result in monetary penalties). Accordingly, there have been a number of studies into agreement in legal contexts, specifically eDiscovery [3, 6, 7, 20, 21, 23, 24]. Oard and Webber [11] found that Cohen's κ ranged from 0.31 to 0.59 in electronic discovery contexts and that the variability in agreement was generally smaller among assessors than topics. Outside of electronic discovery context, we know of no other studies in legal contexts that examines assessor agreement.

3 STUDY METHODOLOGY

The high-level procedure for our study was for a participant to annotate 50 documents for 5 common due diligence topics ("Change of Control", "Assignment", "Indemnity", "Exclusivity", and "Most Favoured Nation"). We note that participants were restricted to one of two assessing strategies but since our focus is on participants' due diligence experience, we omit details and analysis of the different strategies for future work. Participants were asked not to go back and alter any previous annotations they had made. The annotation process was conducted using in our in-house platform [15]. [1]

Participants were solicited from our in-house team of annotators and from other lawyers in our organization. In total, we had 9 respondents, four of which had previously had substantial experience with due diligence review. All annotators were at least familiar with the topics and generally found them easy according to self-report on a 5-point Likert scale. Additionally, one of the authors conducted a full-scale review of the document set to provide a baseline for performance and ensure that the documents were representative of the task. This annotator is referred to as the "Gold" annotator while all others are identified in aggregate as Lo(w) experience or Hi(gh) experience annotators.

While all participants were familiar with the topics they were asked to annotate, we provided additional fine-grained details about what should and should not be annotated in the course of their review. An example for the "Indemnity" topic can be seen in Figure 1. This reflects the real-world scenario where a senior lawyer will dictate what is and is not relevant for a particular review. We note that there has been evidence that instruction length can have non-trivial effects on assessing behaviour [1]. Such trends do not seem to hold in legal domains as more and less detailed instructions do not appear to produce substantial differences in assessor behaviour [24]. As none of our topic descriptions are particularly long nor extremely detailed, we do not believe this had a substantial effect on our results.



5) Indemnity

- include positive obligations to indemnify/hold harmless as well as indemnification procedures
- generally you will be highlighting entire indemnity sections
- exclude separate survival provisions referencing indemnities
- exclude separate defined terms relating to indemnities (but include if embedded in the provision)

Figure 1: Example of the additional instructions supplied to annotators as part of the study. They were expected to be otherwise familiar with the topic.

	T1	T2	T3	T4	T5
Gold	77 (490)	29 (353)	49 (533)	81 (1404)	9 (352)
A	63 (554)	39 (361)	45 (504)	67 (1662)	6 (395)
B	60 (521)	33 (521)	35 (420)	53 (1879)	11 (360)
C	75 (491)	56 (282)	47 (470)	72 (1528)	11 (322)
D	59 (574)	33 (498)	36 (530)	52 (2020)	6 (391)
E	71 (532)	34 (522)	38 (639)	37 (1826)	12 (490)
F	69 (520)	34 (464)	46 (517)	83 (1476)	9 (409)
G	64 (555)	30 (514)	40 (396)	70 (1639)	6 (448)
H	55 (515)	38 (619)	42 (622)	79 (1209)	9 (538)
I	74 (497)	41 (420)	35 (348)	69 (1592)	6 (452)

Table 1: Number of annotations and the average number of characters in an annotation for each of the five topics for the annotation task.

As can be seen in Table 1, by and large, all of the annotators identified a similar number of instances of each topic. There does appear to be a trend to either have more, shorter annotations or fewer, longer annotations. Some participants tend to follow this pattern quite rigidly (e.g., D), while others (e.g., E) fluctuate from topic to topic.

3.1 Evaluation Measures

Following Voorhees [18], we report Recall, Precision, and Overlap for pairwise comparisons between annotators, where we switch off the roles as primary (gold standard) and secondary annotator. We note that we only compare our Gold annotator against every one else (i.e., the Gold annotator is only ever used as the primary annotator) since any other comparisons may not be valid as Gold had unrestricted review capabilities. We also report Cohen's κ [4, 10] as it is a long standing measure of inter-annotator agreement. Following Landis and Koch [9], we consider any $0.6 < \kappa < 0.8$ to indicate substantial agreement. For these basic measures, we consider binary relevance to be a function of simple annotation overlap (i.e., any overlap indicates a true positive).

Binary overlap is not particularly nuanced, as a small annotation overlapping a large annotation does not account for the parts that do not overlap. To account for these nuances, we borrow from plagiarism detection [12] and consider macro-averaged Recall and

[1] The data for this study is being released under a restricted usage agreement available at science.kirasystems.com.

Pri.	Sec.	Recall	Precision	Cohen's κ	Overlap
Gold	Lo	0.75 (0.03)	0.84 (0.08)	0.67 (0.05)	0.61 (0.06)
Gold	Hi	0.71 (0.03)	0.82 (0.07)	0.67 (0.05)	0.61 (0.05)
Lo	Lo	0.81 (0.04)	0.93 (0.02)	0.74 (0.04)	0.68 (0.04)
Hi	Lo	0.89 (0.02)	0.81 (0.02)	0.76 (0.02)	0.70 (0.02)
Hi	Hi	0.82 (0.03)	0.87 (0.04)	0.77 (0.04)	0.70 (0.05)

Table 2: Mean and standard deviation of various binary relevance measures (i.e., span overlaps or not). The top lines represent when our Gold standard annotator is used to evaluate all other annotators. The remainder specify when each experience level is used as the primary annotator and evaluates another annotator as secondary assessor.

Precision, gRecall and gPrecision, which account for the amount of overlap (or lack thereof) between annotations. Similarly, we can extend simple Overlap to gOverlap, such that we divide the length of the intersection of annotated text by the length of the union of all annotated text.

For the sake of brevity, we report measures with respect to meaningful combinations of Gold, Hi(gh) experience, and Lo(w) experience assessors. Since we are averaging over assessor pairs and topics, we use a fixed-effects model [5] to appropriately weight differences in mean and variance across the pairs of assessors and topics. For the purposes of our evaluation, the Pri(mary) assessor is the gold standard for evaluation and the Sec(ondary) assessor is the one being evaluated.[2]

4 BINARY AGREEMENT

In Table 2, we see that there is substantial agreement among annotators using binary measures, with annotations overlapping more often than not. Our κ and Overlap values are higher than those reported in the literature [3, 16, 18], which may be due to the fact that these are not strictly topical in nature. Namely, "Change of Control" has a precise legal meaning; whereas, identifying that a document is about "black bear attacks," or "fantasy football gambling at Enron" may be more open to interpretation. It has been argued [7, 11] that, in the eDiscovery context, it is often human error, rather than differences of opinion, that causes variations in assessments. It is worth noting that Wakeling et al. [19] found similar levels of agreement when examining agreement in real-world search topics (i.e., issued by real users rather than constructed for a particular task), which indicates that agreement we have observed may be due to real-world nature of our topics.

The participants achieve Precision on par with previous studies [6, 18] but Recall is higher than that reported by Voorhees and others [16, 18]. The disparity between what our study and other legal contexts have observed [3, 16] stems from the task itself. Namely, eDiscovery requires retrieval from a much larger document collection and so increases difficulty in achieving high Recall and Precision. While our values may be in-line with previous results, they are still quite low from a practical standpoint. Being overly inclusive means burdening a costly senior lawyer with additional review or missing a potentially important but cleverly hidden piece of information. Both issues could have consequences for the lawyer who led the review.

Pri.	Sec.	gRecall	gPrecision	gOverlap
Gold	Lo	0.71 (0.04)	0.73 (0.11)	0.61 (0.08)
Gold	Hi	0.69 (0.04)	0.73 (0.11)	0.66 (0.08)
Lo	Lo	0.75 (0.05)	0.82 (0.04)	0.64 (0.06)
Hi	Lo	0.79 (0.02)	0.78 (0.02)	0.69 (0.04)
Hi	Hi	0.77 (0.04)	0.80 (0.04)	0.70 (0.07)

Table 3: Mean and standard deviation when we take into account the granularity of the annotated spans. gRecall and gPrecision are the macro-averaged versions by Potthast et al. [12] that take into account overlap length of annotations.

Similar to Bailey et al. [2], we do find some assessing differences between our more and less experienced annotators. When we compare High and Low experience assessors, there is higher Recall at the expense of Precision, which we may expect given that highly experienced lawyers would have more refined conceptions of the topics than less experienced ones. Interestingly, when compared to the Gold assessor, there do not appear to be meaningful differences between the levels of experience but this may be due to the study design and time constraints. When comparing agreement among the different experience levels, we do not observe a meaningful difference in κ or Overlap. This appears to support the finding of Kinney et al. [8] that background experience may not have substantive impact on assessor agreement.

5 GRANULAR AGREEMENT

The more granular measures of agreement are our mechanism for examining the nuances in the amount of material annotated by annotators. As seen in Table 3, gRecall and gPrecision are noticeably decreased from their equivalents in Table 2. This appears to indicate that even though users agree on the general area of relevant information in documents (as corroborated by the discussion in the previous section), they do not agree on the amount or exact context. Further evidence of this can be seen by the generally same levels of gOverlap and Overlap but higher variance. It is reasonable to expect this would only occur when the granularity of the annotation is the main source of disagreement rather than general area.

There is also a general increase in variation for the granular measures when compared to their coarse counterparts. From this we may infer that the granular measures may provide us with a more in-depth picture into how participants agree with each other. In particular, if we care about the amount of information annotated then there will be decidedly more disagreement between some assessors than Table 2 would lead us to believe. This is of particular importance since, in the real-world, these annotations would be used to generate a report on whether or not a transaction (e.g., merger, acquisition) is worthwhile. For example, if a senior lawyer expects concise annotations to expedite report generation then we would suggest that the annotating lawyer likely target that goal rather than try to capture surrounding context. While the surrounding context may be helpful to the more junior lawyer, it would likely distract the senior lawyer from their task.

[2] Note that these measures are symmetric and switching who is the Primary only flips Recall and Precision.

	T1		T2		T3		T4		T5	
	P	R	P	R	P	R	P	R	P	R
Gold	0.82	0.83	0.78	0.60	0.72	0.66	0.77	0.83	1.00	0.33
Lo	0.83	0.93	0.80	0.78	0.68	0.50	0.78	0.90	0.85	0.21
Hi	0.81	0.94	0.80	0.78	0.63	0.44	0.76	0.89	0.96	0.29

Table 4: Average P(recision) and R(ecall) for Gold, Low experience, and High experience models trained by 5-fold cross validation evaluated against the training user.

In spite of these differences, the large scale trends appear to be the same as the previous section. Accordingly, the binary measures may be a suitable measure for assessing general assessor quality. The more granular measures may be more useful when attempting to match assessing styles between lawyers to mitigate any subsequent disagreement on the amount of material highlighted.

6 LEARNING RESULTS

In our study, it was feasible, if time consuming, to review the 50 documents for 5 fields. In practice, where there are thousands of contracts, the costs of having lawyers review documents is not feasible. Accordingly, we can take the annotations provided by our participants and train a user-specific machine learning model for each topic to identify instances in unseen documents. Following Roegiest et al. [14], we trained a Conditional Random Field ("CRF") for each participant and each topic. We generated each model's Precision and Recall with respect to its training assessor and report the aggregates across experience level in Figure 4.[3] The models trained by highly experienced lawyers do not appear to be substantially more accurate than those trained by the less experienced ones. Interestingly, our Gold assessor does not appear to produce a substantially superior model and goes to show that these models are imperfect copies of the underlying assessor. With more examples, this may be partially mitigated but how much is unknown.

We then ran those CRFs on 20 previously unseen documents. Figure 5 depicts κ, Overlap, and gOverlap when comparing model annotations across experience levels. When compared to Figures 2 and 3, we find that generally agreement is in accord with the annotator-level agreement. Though we do find increased variation compared to previous results. What we might infer from this is that, on average, the models produce roughly the same annotations as their training users might have. But the models have the potential to magnify the differences in their trainer's conception of relevance. To this end, two participants were able to annotate the 20 documents themselves and they achieved average κ values of 0.56 and 0.67 with respect to their models, lower than we might desire but not surprising given the above. Such a result further reinforces the idea that the models generated are imperfect copies.

Ideally, we would like the models trained on annotations to be generalizable such that another user would find them useful. While the models achieve what is considered substantial agreement [9], we are still not quite there in light of the results above. Part of the reason for this may be due to an insufficient amount of training data, but may also result from the restriction that participants could not go back and refine their opinion on previously judged documents and, thus, fix errors. All participants stated that this deviated from their general reviewing routine. Further exploration is necessary to determine whether a less restrictive annotation strategy would produce a better model and higher agreement. Alternatively, it may be the case, as Roegiest et al. [13] suggest, that combining annotations from multiple users could yield a more effective model and we leave this for future work.

Pri.	Sec.	Cohen's κ	Overlap	gOverlap
Gold	Lo	0.64 (0.08)	0.59 (0.08)	0.65 (0.10)
Gold	Hi	0.69 (0.09)	0.64 (0.10)	0.67 (0.11)
Lo	Lo	0.70 (0.04)	0.64 (0.04)	0.70 (0.05)
Hi	Lo	0.74 (0.02)	0.68 (0.03)	0.71 (0.03)
Hi	Hi	0.72 (0.05)	0.66 (0.07)	0.69 (0.07)

Table 5: Mean and standard deviation of the agreement measures between the results of applying machine learning models trained by each annotator's assessments on 20 previously unseen documents.

7 CONCLUSION

We have presented a study of assessor agreement when annotating passages in documents for the due diligence context. We have found that legal professionals tend to agree with each other more than assessors did in previous studies. However, this agreement diminishes when we account for the length of the annotated material. Indeed, there is more variability in agreement when accounting for amount than would be suggested by presence of any annotation overlap. In spite of this, the variability observed tends to be less than that found in other legal contexts indicating that perhaps the topics are sufficiently well understood. This suggests that the differences in assessing behaviour are what remain. While some small differences in agreement exist between more and less experienced users, there is no strong evidence that one is inherently better than the other. We place the caveat that our results form an "optimistic" estimate of agreement as the participants are familiar with internal definitions of these topics. Accordingly, we might reasonably expect more disagreement between professionals from different institutions.

When using the annotations rendered by users to train a machine learning model, we find that the resulting models exhibit similar levels of agreement as the constituent users. This has interesting implications as it means that presenting the annotations of one user's model to another will likely yield similar disagreement to presenting the trainer's annotations. Such an implication means that we need to be careful and ensure that when training such models, the annotations should target as close as possible to a "reasonable" interpretation of relevance.

ACKNOWLEDGMENTS

The authors would like to thank our in-house team of legal professionals, whose participation made this study possible. The authors would also like to thank Karen Anello for finding and organizing the documents that were used in this study.

[3] The performance of T5 ("Most Favoured Nation") is low due to the relatively few examples in the 50 study documents.

REFERENCES

[1] Aiman L. Al-Harbi and Mark D. Smucker. 2014. A Qualitative Exploration of Secondary Assessor Relevance Judging Behavior *(IIiX '14)*.

[2] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. 2008. Relevance Assessment: Are Judges Exchangeable and Does It Matter *(SIGIR '08)*.

[3] Thomas Barnett and Svetlana Godjevac. 2011. Faster, better, cheaper legal document review, pipe dream or reality?. In *Proc. ICAIL DESI IV Workshop*.

[4] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960).

[5] Gordon V. Cormack and Thomas R. Lynam. 2006. Statistical Precision of Information Retrieval Evaluation *(SIGIR '06)*.

[6] Maura R Grossman and Gordon V Cormack. 2010. Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review. *Rich. JL & Tech.* 17 (2010).

[7] Maura R. Grossman and Gordon V. Cormack. 2011. Inconsistent Assessment of Responsiveness in E-Discovery: Difference of Opinion or Human Error?. In *Proc. ICAIL DESI IV Workshop*.

[8] Kenneth A. Kinney, Scott B. Huffman, and Juting Zhai. 2008. How Evaluator Domain Expertise Affects Search Result Relevance Judgments *(CIKM '08)*.

[9] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33, 1 (1977).

[10] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* Cambridge University Press.

[11] Douglas W Oard and William Webber. 2013. Information retrieval for e-discovery. *Foundations and Trends® in Information Retrieval* 7, 2–3 (2013).

[12] Martin Potthast, Benno Stein, Alberto Barrón-Cedeño, and Paolo Rosso. 2010. An evaluation framework for plagiarism detection. In *COLING '10*.

[13] Adam Roegiest, Gordon V. Cormack, Charles L.A. Clarke, and Maura R. Grossman. 2015. Impact of Surrogate Assessments on High-Recall Retrieval *(SIGIR '15)*.

[14] Adam Roegiest, Alexander K. Hudek, and Anne McNulty. 2018. A Dataset and an Examination of Identifying Passages for Due Diligence. In *Proc. SIGIR 2018*.

[15] Adam Roegiest and Winter Wei. 2018. Redesigning a Document Viewer for Legal Documents. In *Proc. CHIIR 2018*.

[16] Herbert L. Roitblat, Anne Kershaw, and Patrick Oot. 2010. Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review. *J. Am. Soc. Inf. Sci. Tec.* 61, 1 (2010).

[17] Andrew Trotman and Dylan Jenkinson. 2007. IR evaluation using multiple assessors per topic. In *Proc. ADCS '07*.

[18] Ellen M Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information processing & management* 36, 5 (2000).

[19] Simon Wakeling, Martin Halvey, Robert Villa, and Laura Hasler. 2016. A Comparison of Primary and Secondary Relevance Judgements for Real-Life Topics *(CHIIR '16)*.

[20] Jianqiang Wang and Dagobert Soergel. 2010. A user study of relevance judgments for E-Discovery. *Proc. Am. Soc. Info. Sci. Tech.* 47, 1 (2010).

[21] William Webber. 2011. Re-examining the effectiveness of manual review. In *Proc. SIGIR 2011 SIRE Workshop*.

[22] William Webber, Praveen Chandar, and Ben Carterette. 2012. Alternative Assessor Disagreement and Retrieval Depth *(CIKM '12)*.

[23] William Webber and Jeremy Pickens. 2013. Assessor Disagreement and Text Classifier Accuracy *(SIGIR '13)*.

[24] William Webber, Bryan Toth, and Marjorie Desamito. 2012. Effect of Written Instructions on Assessor Agreement *(SIGIR '12)*.