A Dataset and an Examination of Identifying Passages for Due Diligence

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

ABSTRACT

We present and formalize the due diligence problem, where lawyers extract data from legal documents to assess risk in a potential merger or acquisition, as an information retrieval task. Furthermore, we describe the creation and annotation of a document collection for the due diligence problem that will foster research in this area. This dataset comprises 50 topics over 4,412 documents and ~15 million sentences and is a subset of our own internal training data.

Using this dataset, we present what we have found to be the state of the art for information extraction in the due diligence problem. In particular, we find that when treating documents as sequences of labelled and unlabelled sentences, Conditional Random Fields significantly and substantially outperform other techniques for sequence-based (Hidden Markov Models) and non-sequence based machine learning (logistic regression). Included in this is an analysis of what we perceive to be the major failure cases when extraction is performed based upon sentence labels.

ACM Reference Format:
Adam Roegiest, Alexander K. Hudek, and Anne McNulty. 2018. A Dataset and an Examination of Identifying Passages for Due Diligence. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3209978.3210015

1 INTRODUCTION

The goal of due diligence in mergers and acquisitions ("M&A") law [14, 30, 31] is to identify all passages in a set of legal documents, often contracts, that would pose major liability or risk should the transaction occur. While this process, like other legal retrieval tasks [8, 35, 38], has historically been a manual one, recent due diligence blunders [32], in particular HP's estimated $8B loss after acquiring Autonomy [5], have prompted an increased desire for solutions using automated techniques, and correspondingly, a large number of software providers delivering such solutions [1].

In due diligence tasks, lawyers seek to find a set of standard passage-types called provisions, that typically correspond to an increased risk in a transaction. These range from what happens to a contract when one party is acquired ("change of control"), to what parties are involved in the contracts,[1] to what happens when a contract is terminated early. One might naively think, as we did, that this ought to be an easily solved problem as lawyers will generally write contracts in a similar manner to one another. However, there is a surprising amount of variation in contracts as they typically evolve over the negotiation process and incorporate feedback from the involved parties [20]. This complexity is compounded by differences in how jurisdictions phrase and handle provisions as well as more technical problems, such as errors introduced during digitization. Accordingly, we are interested in finding the best approach to identify these passages across a wide variety of legal documents and jurisdictions.

To determine the best approach, a dataset is needed. To the best of our knowledge, no such publicly available collection exists. A fact that we struggled with when developing our own proprietary due diligence platform. To help correct this situation and foster additional experimentation, this paper describes a subset of our own internal dataset that we are releasing with relevant annotations for academic use. This subset spans 50 topics and includes 4,412 manually annotated legal documents totalling over 15 million sentences, collected from various public sources (Section 3).

Using this dataset, we test several possible approaches to identifying desired provisions for extraction. Inspired Natural Language Processing ("NLP") community [11, 15], we treat documents as sequences of labelled and unlabelled sentences and investigate how different sequence-based (Conditional Random Fields and Hidden Markov Models) methods compare to more traditional methods (e.g., logistic regression). Using sentence-level and annotation-level (see Section 4.3) effectiveness measures, we find that CRFs significantly and substantially outperform the other tested methods.

We present an additional examination of the possible degenerate cases that can arise when treating documents as sequences of sentences. We find that CRFs have the lowest incidence rate among the methods tested for all degenerate cases. Finally, we conclude with a discussion of the limitations of this work and potential avenues of further investigation.

2 THE DUE DILIGENCE PROBLEM

Any merger or acquisition that a company undertakes is done to maximize profit and minimize potential risk. The legal process to ensure that risk is minimized is called due diligence. In a typical due diligence case, a group of junior lawyers will be given a collection of documents, mainly contracts but potentially other materially financial documents from the target (read: to be acquired) company,

[1] This can be important if the acquiring party has non-compete clauses or similar.

SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA

ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210015

and will be told to find all relevant passages with respect to various information needs (e.g., change of control, mentioned parties, contract values, etc) within a window of time. If the documents have been stored electronically (in a virtual data room), they will often use a combination of manual review and keyword-search [14, 30] to first identify the relevant documents and then manually find the desired information in those documents and copy it into a spreadsheet. Otherwise, they are left to review physical documents. Confounding this is the time sensitive nature of diligence reviews, which often means only a small subset of the documents are being examined. Risk and potential profit is projected for the entire portfolio of the target company from this small sample.

Accordingly, we can divide the due diligence problem into two main tasks. The first is to identify all of the relevant passages based upon the various information needs. The second is to use these passages to predict any potential risk to the acquiring company. Note, we do not envision an initial culling step as we believe any reasonable system ought to be able to identify the correct passages without needing any form of culling (i.e., the presence of a relevant passage in a document is a signal it should be included). Though it is important to note that lack of a particular information need in some subset of the document collection may also be an indicator of risk or potentially profit (i.e., a lack of "most favoured nation" provisions, which denotes preferential pricing would generally be considered good). For this work, we concentrate primarily on the first task and leave risk quantification for subsequent investigation.

We would be remiss, however, for not distinguishing the due diligence problem from that of the more established legal retrieval task of electronic discovery ("eDiscovery"), *cf.* [3, 4, 8, 23, 38]. In short, eDiscovery is a phase of civil litigation that requires the defendant to produce all, or nearly all, electronic data that is relevant (and unprivileged) of the case to the plaintiff. The amount of relevant material returned is often required to be reasonable and proportionate to the needs of the case (e.g., a $10k lawsuit does not necessitate $100k of discovery fees). These constraints are generally codified and lawyers held to them (c.f., the Federal Rules of Civil Procedure Rule 26(g) [25]) but not all countries have such rules.

We see the two biggest distinguishing characteristics between eDiscovery and due diligence as: (1) that eDiscovery models tend to be single use since they are tailored to a specific case, while due diligence models ideally apply to any unseen documents (i.e., from the next M&A deal); (2) the granularity of the information returned as due diligence relies on the content of the identified passage. Indeed, these two characteristics will shape how we label data and assess effectiveness of algorithms in Section 4. Accordingly, we believe that while there may be overlap in some aspects of these two problems, the solutions will be substantively different though not necessarily disjoint.

3 DOCUMENT COLLECTION

The document collection used throughout this work is a combination of legal documents (e.g., credit, loan, and facility agreements), topics describing the information that should be extracted (e.g., "produce all passages relating to changes in ownership of company X"), and annotations (i.e., sentence labels) mapping pertinent portions of a document to the relevant topic. Accordingly, a single document may have multiple annotations for a single topic and for multiple topics depending on the document's type and its contents.

As part of Kira Systems' commitment to promoting scientific inquiry and the replicability and repeatability of the experiments described herein, we are disseminating the collection (with annotations) for use in the academic community. Access to the data will require signing an agreement describing the situations in which the data can be used (i.e., academic/non-commercial use). The data usage agreement and instructions are available at https://science.kirasystems.com.

The Kira Systems collection, while the raw documents are publicly available, does comprise of several hundreds of work-hours in its creation. In particular, it is worth noting the data provided and used in this work is exactly what is used in our production system [37] to produce client facing machine learning models. We have not redacted, transformed, simulated, or augmented the data in any shape or form other than those ways described in this paper and in any subsequent errata. To this end, the following subsections provide additional details on the document provenance, the topic development process, and the annotation methodology used to generate the labellings in the dataset.

3.1 Document Provenance

Unsurprisingly, law firms and companies are not particularly willing to share their legal documents with outside sources, at least not without redaction. As a result, finding sufficient amounts of legal documents to train machine learning models can be particularly difficult, especially when one tries to ensure diversity among document types. In the United States and Canada, publicly traded companies are required to file material documents (e.g., contracts, executive-level employment agreements) to public data sources which helps to ameliorate the issue. In the US, this is the Electronic Data Gathering, Analysis, and Retrieval system ("EDGAR"), and, in Canada, the System for Electronic Document Analysis and Retrieval ("SEDAR").

While generally helpful, such sources have their own limitations due to the nature of the filings. For example, non-executive employment agreements or retail real estate leases may also be examined when conducting a due diligence review (or other similar reviews), though these documents are typically not filed under EDGAR/SEDAR. In those circumstances, we have had to search across various sources, none of which are particularly consistent.

While other countries have similar sources to EDGAR and SEDAR, we have not found them to be particularly useful in finding documents for our use cases. This is primarily because financial reporting laws of public companies are different than in Canada and the US. Accordingly, in this dataset and in many of our own models there is a bias towards Canadian and American law. However, we note that this dataset does contain a small set of documents from the UK.

3.2 Topic Development

Our topic development process is generally guided by what our customers want, what we think they will need, and what additional markets we might want to look into (e.g., lease abstraction). This requires a non-trivial amount of educated forecasting, not dissimilar

```
<title>
Accounting Changes Negative Covenant

<description>
When one or more lenders extend credit to a borrower,
it is typical for the lender(s) to require that the
borrower promise to take certain actions and refrain
from taking other actions while the loan is
outstanding. This topic captures covenants of a
borrower (in credit, facility or loan agreements)
not to make any changes to its accounting practices,
change its fiscal year end or change its accounting
reference date while the loan is outstanding.
```

Figure 1: An example of the description provided to users for a particular topic.

to what TREC Real-Time Summarization assessors are required to do [18]. For example, there are many ramifications from the United Kingdom leaving the European Union (i.e., Brexit) as is the introduction of the GDPR, so having models capable of extracting useful information related to these issues provides utility to customers.

What this all amounts to is coordination between large parts of our company (e.g., product, annotation, sales, customer success) to determine what topics can bring the most value. Our efforts are often hampered by a lack of documents exhibiting the desired topic, so we are left to explore other topics or spend effort trying to find example documents should the estimated benefit be sufficient. Ultimately, our end goal is to provide customers with useful and generalized topics and models so that they do not need to train their own bespoke models.

Topics are typically workshopped until we can determine that they are viable (i.e., we have enough examples to produce an acceptable model). Once this is done, we roll-out these models to customers with a short description of what they are intended to find and the types of documents they are used for. Figure 1 presents an example topic description, while the remaining 49 are omitted for brevity.

All the topics provided in this dataset have undergone this process and have corresponding models available for use by end users of our systems. As far as we are aware, such a release of proprietary training information in the legal domain has never been undertaken. We take on such a release with the belief that the benefit to the scientific community outweighs any potential risk.

We note that the topics released in this dataset relate to diligence and knowledge management-related reviews and not what might be construed as "core" due diligence. Namely, we have omitted the release of training data for "Assignment" and "Change of Control" topics as they are an important differentiator from our competitors.

3.3 Annotation Methodology

Prior to annotation, we ingest documents into our system which performs Optical Character Recognition ("OCR"), various pre-processing steps for machine learning features, and metadata generation (e.g., document type classification). Once this is complete the documents are available in our document viewer, described previously by Roegiest and Wei [29], which allows users to annotate documents and review annotations of other users.

	Avg	Min	Max
Documents	305.42	78	592
Documents w/ Example	211.60	24	555
Sentences	1,022,442.32	262,242	2,077,330
Relevant Sentences	1,526.16	118	9,897

Table 1: Average, minimum, and maximum number of documents, documents with at least one example of a topic, sentences, and sentences that overlap with a user annotation across the 50 topics provided in the Kira Systems collection.

Document annotation by our in-house annotators (consisting of law students, contract lawyers, and in-house senior lawyers) consists of the following phases: initial annotation of ~20 documents, quality control with a senior lawyer, refinement of the annotations, additional annotation of 20 to 40 documents,[2] and then an initial training of the system. Once the training is complete, the annotator reviews the false positives and false negatives and refines the existing annotations to reflect any deficiencies in the model. Following this refinement, the annotator has a follow-up review with a senior lawyer to discuss what further modifications need to be made. Such modifications can include adding additional examples, further refinement of the annotations, or, in extreme cases, removal of documents that are skewing the model in negative ways. This process repeats until our internal metrics report an annotation-level precision of 85% and a recall of 90% in many cases, though depending on the exact use of the model we may be more or less strict on that requirement (i.e., not identifying pertinent text may impose significant risk). It is worth noting that we use law students solely as annotators; they are omitted from any training or model refinement.

Finally, each model has a final *in situ* test. We apply the model to a previously held out set of documents and a senior lawyer reviews the automatically generated annotations for correctness. If we find the model to be lacking, we add additional training data or further refine existing highlights. In general, our senior annotators often err on the side of caution and apply a policy of "would a reasonable lawyer consider this relevant to the topic?" Coincidentally, this is not dissimilar from the advice given by Roegiest et al. [28] when discussing evaluation of high-recall retrieval systems.

This annotation process invariably leads to some bias in the resultant annotations. Such bias is arguably unavoidable if we want to produce high-quality models capable of annotating legal documents in desirable ways for end users (i.e., we want to provide training data in the best possible form for the underlying algorithm). However, this is not to say that we have tailored the data to the algorithm. Indeed, we are more likely to add additional examples to encourage the algorithm to identify the text we believe to be important, rather than eliminating examples or changing annotations to get the highest score possible. Doing so would yield inferior models and compromise our core business built on the strength of our provided "out of the box" models that are widely applicable.

Table 1 provides some insight into the composition of our training data once we deem a model to be "good enough" for the 50

[2] Depending on the perceived difficulty associated with a topic, the numbers of reviewed documents may be dramatically higher.

topics we use in this work. In total there are 4,222 unique documents and 15,157,820 total sentences in the dataset and they are primarily all credit related agreements.

While document-level prevalence in this collection is in fact quite high on a per-topic level basis (i.e., two-thirds of documents have a relevant example on average), the more interesting aspect of the dataset is that there are far more unannotated sentences than there are annotated ones. This is a much more useful piece of information than document-level prevalence as our task looks at sentence rather than document-level relevance. Accordingly, in the best case topic, we have a prevalence of 0.7% and 0.01% at worst. Identifying the correct sentences, even if they are obvious, is unlikely to be as straightforward as we might otherwise hope.

4 EXPERIMENTAL METHODOLOGY

Using the aforementioned dataset (Section 3), the experiments in this work follow the Cranfield paradigm [7] to compare the effectiveness across a selection of algorithms and libraries (Section 4) that encompass what we believe to be potential reasonable solutions. To do so, we split the documents into 5 roughly equal folds and run 5-fold cross validation on them where evaluation scores, whose computation is detailed in Section 4.3, are generated in aggregate across test folds. Note that we do not make any explicit attempts to mitigate the extremely low prevalence of relevant sentences. Each learning method has the entirety of each training fold available to them. This is crucial for sequence-based approaches as they rely on contextual information but may hinder other approaches. While we could attempt to re-balance the dataset for non-sequence approaches, investigation into the best approach for this task is worthy of its own investigation and so we leave that for subsequent examination.

As discussed earlier, we split documents into sentences and label each sentence according to our existing annotations as relevant (should be annotated) or not relevant (should *not* be annotated). We discuss the sentence segmentation and feature generation for these experiments in Section 4.2. Finally, all statistical tests are paired t-tests where † denotes $p < 0.01$ and ‡ denotes $p < 0.0001$. We note that these values are uncorrected but Bonferroni correction would still yield significance at the $p < 0.05$ level.

4.1 Methods

In this section we detail the various algorithms and libraries employed in these experiments and our motivations for their use. Before doing so, we must address a method that we *do not* employ which are any keyword/rules-based methods for sentence labelling. Such approaches have been extensively used in ad-hoc retrieval [6, 36, 39] and question-answering tasks [33] but we believe such methods are lacking in the due diligence scenario for several reasons. The first, as seen in the TREC Legal tracks [4, 12, 24], keyword-only approaches often ended up being long, complicated, and thus, fragile (e.g., terms not in the keyword query are ignored), and did not perform overly well to more complicated machine learning techniques. The second, has seen many topics in practice that have not historically performed well with keyword-based approaches. Consider Figure 2, these show four hard to find instances

False Miss	In the event that, at any time, TWC or a TWC Affiliate no longer controls the WCG Group this Agreement shall terminate 90 days from the date control was lost.
False Miss	Either party, immediately upon written notice to the other party, may terminate this Agreement upon the Merger or Bankruptcy of the other party.
False Miss	Owner shall have the right to terminate this Agreement (i) following any failed drug test by the senior officers of Manager and the failure of Manager to promptly take remedial action in connection with such senior officer, (ii) any failed drug test of William W. Warner, (iii) following the occurrence of an Event of Default by Manager that is not cured within the applicable cure period described in Section 15.1, (iv) if Manager is no longer controlled by William W. Warner (including, for the avoidance of doubt, in connection with death or disability), or (v) if William W. Warner is convicted of, or pleads nolo contendere (or a similar plea) to any felony.
False Miss	Without the prior written consent of the Lender, the Borrower shall not: (a) cease to be a wholly-owned Subsidiary of Denison.
False Hit	If CSI adds, deletes, or changes a manufacturer/distributor/supplier, CSI will notify FRESENIUS KABI and FRESENIUS KABI will issue a Change Control request prior to implementing the change.
False Hit	Except to the extent provided in Section 6.10(b) below, Client shall pay Patheon the amounts incurred in implementing a change to the Specifications requested by Client under this Section 6.10(a), as determined in accordance with the Change Control Procedure.

Figure 2: Examples of incorrect hits and misses for a hypothetical keyword-based search for "Change of Control" (e.g., what happens to a contract if one party is acquired) instances that we have trained a machine learning model to correctly identify.

of "change of control" and two passages that would be easy to mislabel as "change of control" (e.g., any query mentioning "control" following "change" within a certain proximity). While we freely admit that these are contrived, but real, examples that we could develop queries to label correctly, we may then incorrectly label some other set of passages. The query would then have to be reformulated to account for those incorrect labels. This process would repeat as new examples are found and the resulting query or set of queries would get larger and more complex and be potentially prone to errors as they get larger. Accordingly, we believe that using such approaches as baselines would yield weak baselines and not be as informative as one might like.

We should note that subsequent keyword-based search over previously identified sentences and passages is a task lawyers can and do undertake. Accordingly, our production system does allow search over such text. Keyword-based retrieval is useful in tackling the due diligence problem but not necessarily for the *core* task of identifying relevant sequences of sentences.

4.1.1 Conditional Random Fields. A Conditional Random Field ("CRF")[15] is a probabilistic graphical model for labelling sequences. It improves upon Hidden Markov Models by being a discriminative rather than a generative classifier and can also consider features

from neighboring sequence positions when evaluating a specific sequence position.

A typical application of a CRF is entity extraction, where text is modeled as a sequence of tokens and the goal is to label each token as entity or non-entity. This works extremely well when the entity you are extracting is relatively short, such as the name of a person or a date. Indeed, CRFs have typically been one of the top contenders for various NLP labelling tasks [11, 34]. Modeling text in this way does not work nearly as well for identifying longer sections of text, such as sentences or paragraphs. Since legal clauses are typically multiple sentences in length, we have not been effectively able to use CRFs to find them using sequences of tokens. Thus, prompting the use of sequences of sentences.

For the experiments in this paper, we use the CRFsuite software [26] which provides a variety of training algorithms and was the most performant of the packages we examined during initial development, and at the time, was the only CRF software that could handle arbitrary numbers of features. In particular, we have found the Passive-Aggressive [9] training algorithm to be very effective and inherently have a slight bias towards recall. Accordingly, we report results with our slightly tuned parameters for CRFsuite that have been effective for us in the past. To provide a baseline, we also report the performance when using the default LBFGS [22] training algorithm, which we have observed to prefer precision. Where disambiguation is necessary **PA** denotes the former training and **LBFGS** the latter.

We used the following parameters during training CRFsuite:

- **PA**: `-a pa -p c=0.1 -p type=2 -p max_iterations=100`
- **LBFGS**: `-p max_iterations=100`

There are two caveats: (1) we restricted CRFsuite to a maximum of 100 iterations because it would otherwise iterate until convergence; (2) the PA parameters result from hyper-parameter tuning.

4.1.2 Additional Sequence Learning. While it would be tempting to consider a pure Hidden Markov Model for comparison to CRFs in these experiments, we instead opt to use SVM$^{\text{hmm}}$[3] and its generalization [34] of Altun et al.'s combination of HMMs and SVMs [2]. This change is primarily due to the general superiority of CRFs over HMMs, discussed above, and the generally competitive nature of SVM$^{\text{hmm}}$ to CRFs for a variety of tasks. The main benefit to this approach is that it facilitates the learning of non-linear discriminant functions while also overcoming limitations of HMMs, such as the inability to deal with certain types of features. The essence of this approach is discriminative models are trained that are isomorphic to an equivalent Hidden Markov Model.

For the purposes of our experiments in this paper, we were largely guided by the settings of SVM$^{\text{hmm}}$ for entity recognition in [34], advice on the software's homepage, and some minor tuning to get acceptable runtime for these experiments. As a result, we used the parameters `-c 1 -e 1` 'c' denotes the traditional slack trade-off variable and 'e' the tolerance of the solution. Smaller 'e' values increase training time (i.e., convergence) and memory usage non-trivially from our anecdotal experience.

4.1.3 Linear Classification. Given that we are treating each sentence as its own object for labelling, we can treat the task as one of binary classification. Admittedly, this is perhaps somewhat of a stretch, but does provide a more meaningful baseline and reflects one of our first investigations into methods for performing this task. Accordingly, we make use of the Vowpal Wabbit ("VW") machine learning package [17] to train a linear classifier with logistic loss. This embodies a logistic regression classifier and would allow us, if we wished to do so, to generate a probability for each sentence detailing whether or not it should be labelled. The benefit to this approach is that we could then fine-tune thresholds for whether or not to label a sentence as "should be annotated." For experimental simplicity, we only report and use the binary classifications as made by VW's `--binary` option in prediction mode.

We run VW with two sets of features, the tuned ones described in the subsequent section, and the auto-generated features using VW's featurization functionality, `--ngram 2 -b 24`, when supplied with the tokenized sentence text. In doing so, this creates unigram and bigram features of the tokens and hashes them into a 2^{21} dimensional space. This approach offers a controlled baseline for featurization that is easily replicable. Accordingly, our training scripts use the following parameters:

- **Tuned**: `--holdout_off --loss_function logistic --passes 50`
- **Sent**: `--holdout_off --ngram 2 -b 24 --loss_function logistic --passes 50`

4.2 Feature Engineering

To create sequences of sentences from documents, we must segment documents in some fashion. While there are a variety of approaches, we internally use and have used for this work an in-house implementation of Kiss and Strunk's "punkt" algorithm [13]. Our motivation to use the "punkt" algorithm is that it allows us to achieve near state of the art performance while allowing us to train the segmentation model to be sensitive to legal documents, which have many abbreviations and shorthands that caused issues during trials with other solutions. Our sentence segmentation model is trained on approximately 1M documents pulled from the EDGAR repository.

As part of our own internal testing and tuning, we examined a variety of possible ways to featurize each sentence, including word token n-grams, stemming, part-of-speech tagging, and case normalization. We found that case normalization and stemming did not appear to help the labelling task and in some cases harmed performance. Part-of-speech tagging was also found not to be particularly beneficial.

On the other hand, we have found that binary (i.e., present or not) token n-grams works very well for this task. Including n-gram frequency did not help, which we posit is due to sentence length being relatively short. Additionally, more typical sequence labelling features generated from neighbouring sentences was not a help. All said, the resulting features that we have used are similar to those used by others for related sequence-oriented tasks [11].

Additionally, we introduce new *word vector* token bigram/trigram features that are helpful when training sets are small. To create these features, we trained a vector representation using `word2vec` [21] on EDGAR documents. These vectors are then clustered using k-means clustering. For each word in a sentence, we create the associated

[3] https://www.cs.cornell.edu/people/tj/svm_light/svm_hmm.html

bigram/trigram based on each word's associated cluster identifier. Due to space limitations, we do not report evaluation on all these combinations but plan to do so at a later time.

Finally, for our linear classification runs, we also conduct separate runs resulting from feeding the tokenized sentences into the Vowpal Wabbit library and allow it to create its own sequence of features using the built-in n-gram tokenizer set to bigrams. Where necessary, **Tuned** denotes the use of our in-house tuned features and **Sent** denotes VW created features.

4.3 Evaluation Measures

While there are many possible ways in which we could choose to evaluate the performance of the algorithms tested, we focus on precision, recall, and F1 as this reflects what our users see when they train their own models. The reasoning for doing so is quite straightforward, the people training these models are not technical experts and they require simple measures they can easily explain to other non-technical practitioners. Furthermore, these three measures have long been used in legal retrieval and so practitioners are more likely to have some exposure to them. Although, we do acknowledge that we are actively investigating ways to make more nuanced measures palatable to our users.

That being said, the way we measure precision, recall, and F1 is not necessarily common. Namely, we make distinction between sentence-level (i.e., is the sentence labelled correctly?) versus annotation-level (i.e., does a sequence of predicted relevant sentences overlap with a human annotation?). Sentence-level measurement, treating each sentence as its own document, corresponds directly to how IR researchers think of these measures.

Annotation-level measurement, on the other hand, is slightly more nuanced. In essence, we heuristically group contiguous sentences that are predicted as "should be annotated" as a single annotation. Thus, a true positive occurs when a gold standard annotation overlaps with a predicted annotation. False positives and false negatives occur in the obvious way.

Figure 3 depicts a hypothetical scenario where we can discuss differences between the two types of measurement. In the depiction, a user has annotated the entire section of text (purple *and* pink) but the trained algorithm has only identified a portion of the annotation. Under sentence-level measurement, the machine would receive precision and recall scores less than 1, but under annotation-level measurement, the machine would get perfect scores. This distinction is important and so we report both levels of measurement in this work.

Notwithstanding the fact that annotation-level measurements do elide some important information, we have found that users are not generally receptive to the level of detail in per-sentence scoring as it does not correspond well with their mental model. Annotation-level measures are a convenient compromise that, in our experience, generally correspond to equivalently good performance on sentence-level measures.

That being said, annotation-level measurements do suffer from degenerate cases. Consider the naive retrieval system that marks every sentence as "should be annotated." This system, under both schemes, has a recall of 1 (assuming at least 1 gold standard annotation) but has a precision near 0 and near 1 for sentence-level and annotation-level measurement, respectively. Accordingly, we are actively investigating additional ways to present the potential quality of trained models to our users (e.g., measuring variance in the resultant model's performance) in an easy to understand way.

Method	Recall	Precision	F1
CRFsuite (PA)	**0.85** [0.83,0.88]	0.92 [0.91,0.93]	**0.88** [0.87,0.90]
CRFsuite (LBFGS)	0.80† [0.77,0.83]	**0.94**‡ [0.93,0.95]	0.86 [0.84,0.88]
SVM^{hmm}	0.69‡ [0.64,0.74]	0.93 [0.89,0.96]	0.78‡ [0.74,0.82]
VW (Tuned)	0.62‡ [0.58,0.65]	0.92 [0.91,0.94]	0.74‡ [0.71,0.76]
VW (Sent)	0.65‡ [0.62,0.68]	0.90† [0.89,0.92]	0.75‡ [0.72,0.78]

Table 2: Sentence-level recall, precision, and F1 for all evaluated methods. 95% confidence intervals are presented for all measures where appropriate. † represents a $p < 0.01$ and ‡ a $p < 0.0001$ where all differences are computed with CRFSuite (PA) as the baseline.

5 SENTENCE-LEVEL RESULTS

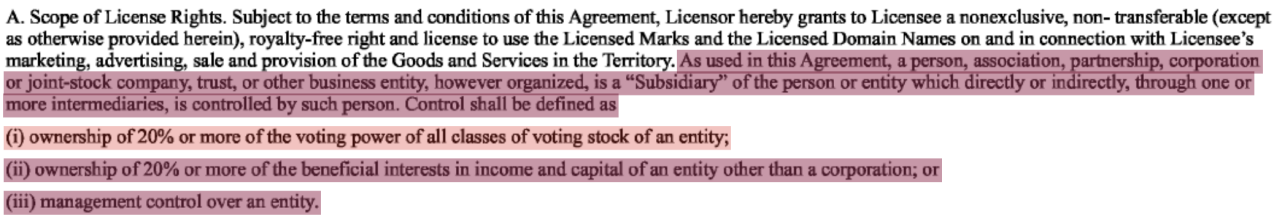
Table 2 summarizes our findings for sentence-level precision, recall, and F1. By and large, our baseline (CRFsuite (PA)) significantly outperforms all other methods with respect to recall and is either superior to, or competitive with, all other methods for precision and F1. This reaffirms our choice to use CRFs (and CRFsuite) in our production environment, as ensuring we have high-recall is often more important to our users than achieving the highest possible precision at the expense of recall.

Confirming earlier remarks, LBFGS appears to prefer precision with PA preferring recall. Even though these differences are not particularly large, they are still significant, and would likely have non-trivial impact in a due diligence case as the lower recall, in the case of LBFGS, could cause an important sentence to be overlooked.

The reasonable performance of Vowpal Wabbit, irrespective of features, indicates that treating this as a binary classification task is a feasible strategy. Though, as will be discussed in Section 6, the low recall will likely have undesirable behaviour at the annotation-level. It is possible that hyper-parameter tuning (to account for the very large class imbalance) would yield increased retrieval effectiveness. However, it is not clear that such hyper-parameter tuning would be fruitful, as the same argument could be applied to the CRFsuite and SVM^{hmm} results.

It is also interesting that our tuned features do not appear to make all that much difference between the VW runs. We might posit that the increased precision for the tuned features results from the word vector n-grams accounting for semantic and syntactic similarities. The improved recall for VW's featurization is less clear but may be benefiting from hits on commonly occurring tokens in the relevant class.

In spite of our relatively simple hyper-parameter tuning, SVM^{hmm} performs surprisingly well with a competitive precision achieved, but at the cost of reduced recall. The naïve solution is to decrease the 'e' value to increase precision of the result. However, this comes with increases in training time and memory usage. There is an additional caveat on the software's webpage that decreasing the parameter below 0.5 does not not typically yield increases in accuracy. Accordingly, we are not convinced that the naïve solution is a useful path as training time and training memory usage are critically important in practical applications. Indeed, even with our


A. Scope of License Rights. Subject to the terms and conditions of this Agreement, Licensor hereby grants to Licensee a nonexclusive, non- transferable (except as otherwise provided herein), royalty-free right and license to use the Licensed Marks and the Licensed Domain Names on and in connection with Licensee's marketing, advertising, sale and provision of the Goods and Services in the Territory. As used in this Agreement, a person, association, partnership, corporation or joint-stock company, trust, or other business entity, however organized, is a "Subsidiary" of the person or entity which directly or indirectly, through one or more intermediaries, is controlled by such person. Control shall be defined as

(i) ownership of 20% or more of the voting power of all classes of voting stock of an entity;

(ii) ownership of 20% or more of the beneficial interests in income and capital of an entity other than a corporation; or

(iii) management control over an entity.

Figure 3: Hypothetical example of machine learning model annotations (dark purple) overlapping with most of a user's annotation (pink) of the entire relevant portion.

Method	Recall	Precision	F1
CRFsuite (PA)	**0.94** [0.94,0.95]	0.92 [0.90,0.93]	**0.93** [0.92,0.94]
CRFsuite (LBFGS)	0.85‡ [0.83,0.88]	**0.97**‡ [0.96,0.98]	0.90† [0.89,0.92]
SVM^{hmm}	0.84‡ [0.81,0.88]	0.92 [0.88,0.95]	0.88† [0.84,0.91]
VW (Tuned)	0.83‡ [0.81,0.86]	0.84‡ [0.80,0.87]	0.83‡ [0.80,0.85]
VW (Sent)	0.88‡ [0.86,0.90]	0.79‡ [0.75,0.83]	0.82‡ [0.79,0.85]

Table 3: Annotation-level recall, precision, and F1 for all evaluated methods. 95% confidence intervals are presented for all measures where appropriate. † represents a $p < 0.01$ and ‡ a $p < 0.0001$ where all differences are computed with CRFSuite (PA) as the baseline.

	Avg	Min	Max
CRFsuite (PA)	0.13	0.03	0.29
CRFsuite (LBFGS)	0.11	0.02	0.29
VW (Tuned)	0.39	0.03	0.85
VW (Sent)	0.40	0.03	0.83
SVM^{hmm}	0.24	0.00	0.44

Table 4: The average, minimum, and maximum fraction of gold standard annotations for which the listed methods only predict partial annotations across the 50 topics.

simple parameters, we ran out of memory when training several topics in parallel on the same machine with 30GB of memory and were forced to train them individually.

6 ANNOTATION-LEVEL RESULTS

The annotation-level evaluation (Table 3) are in accord with the sentence-level results. Primarily, CRFSuite (PA) does appear to be more effective, in general, than the other methods examined. Though we note that the bias for LBFGS to prefer precision at the expense of recall is substantively more pronounced with annotation-level measures. In essence, CRFsuite (LBFGS) and SVM^{hmm} both produce models that are precise in their sequence labelling but miss substantively more documents than CRFsuite (PA), which is not a desirable trade-off when solving the due diligence problem.

As the VW runs show, annotation-level measures can be both more and less forgiving of mistakes made at the sentence level. Recall increases here because the classifier needs to only hit part of the annotation to be counted as a true positive, and since there are fewer true positives overall, recall goes up. A similar case occurs for precision, where a single incorrect sentence label is held to be on par with multiple sequential correct labels (i.e., the reduced numbers of true positives magnifies mistakes).

It is worth noting that doing better at the sentence-level does not always correspond to improved performance at the annotation-level. For example, CRFsuite (LBFGS) and SVM^{hmm} both achieve similar levels of recall to the VW runs at annotation-level but were more effective at the sentence-level, though SVM^{hmm}'s improvement is only marginally better. We do not, however, consider this a flaw in annotation-level measures as the best performing run, CRFsuite (PA), is fairly apparent under both evaluation schemes. Especially since we believe that false positives (i.e., lower precision), while time consuming to deal with, are preferable to false negatives which increase the chance of missing vital information.

We must also point out, as neither Table 2 nor 3 do it justice, that SVM^{hmm} is highly variable in its effectiveness. On several topics, it is very competitive with CRFsuite (PA) and on others it is not. However, there is one topic where it completely fails to mark any sentence as relevant. While we readily acknowledge that we likely have not chosen the best parameters, having this kind of degradation is not useful for a system to have. Indeed, we can easily imagine a scenario where, after optimizing hyper-parameters on some set of topics, a new topic is introduced and SVM^{hmm} fails to perform. The behaviour we have seen here indicates that such behaviour is possible. While this is not bad in theory, in an actual production system where users are attempting to train models, having to do a sweep of hyper-parameters is going to slow training down and yield less than happy customers because of that slowdown.

As we discussed in Section 4.3, there is the potential for a learning algorithm to produce sentence labellings that we consider to be failure case. One such case is partial coverage, where the gold annotation is only partially covered by the predicted labels, and the second is when there is gapped coverage (*cf.* Figure 3). The second case, while a subset of the first, is dramatically less user friendly and one that would lead to a poor user experience if it were exceedingly common.

In Table 4, we show that partial annotation occurs somewhat frequently regardless of the method employed and, in some cases, can be quite disastrous. This is true of VW, which treats each sentence independently of the rest, it only labels the sentences it thinks are most likely to be relevant to the topic and not all sentences that ought to be annotated. The other sentences might not have the necessary features to flag it as being important in isolation.

The CRF approaches do not appear to suffer from this problem as badly, though still more than we might like to see. The why behind this is not clear but we also have not had customer complaints about this being a commonly occurring issue. One potential contributing factor is that our annotators may have been over-inclusive in some fraction of their annotations (in an attempt to cover all possible relevant information). This stems from practical experience where including additional context will sometimes, but not always, help the underlying models discern the actual useful content. Exploring

	Avg	Min	Max
CRFsuite (PA)	0.02	0.00	0.10
CRFsuite (LBFGS)	0.01	0.00	0.06
VM (Tuned)	0.23	0.00	0.73
VM (Sent)	0.24	0.01	0.73
SVM^{hmm}	0.06	0.00	0.14

Table 5: The average, minimum, and maximum fraction of gold standard annotations for which the listed methods only predict gap producing annotations (i.e., multiple predicted spans with partial coverage of a gold annotation) across the 50 topics.

the prevalence of this type of behaviour and its influence on the CRFs is under investigation.

In spite of its generally reasonable performance, SVM^{hmm} partially highlights substantially more gold annotations than the CRF approaches. While we might anticipate this from its lower sentence-level recall, such an outcome is not desirable and may provide further evidence that this approach requires additional thought. Note, the minimum of 0 for SVM^{hmm} corresponds to a complete failure to predict any relevant label for a topic and so is not representative of "best case" behaviour.

Table 5 depicts the occurrence of multiple annotations on a single gold annotation. Perhaps a little unsurprising, the CRFsuite and SVM^{hmm} solutions do not appear to have high incidence rates for this case. Though the worst case scenarios are still not ideal with LBFGS outperforming due to the increased precision (i.e., when it works, it works). VW's performance, on the other hand, is simply bad. If between a fifth to a quarter of the annotations have gaps, users would definitely begin to wonder what is wrong with the underlying model. But like above, the VW rates are likely a side-effect of them treating this as a binary classification task which means that some sentences will not look relevant in isolation.

The performance of SVM^{hmm} here is also very interesting given the dismissal number of partial annotations it makes. It would appear to be the case, when SVM^{hmm} detects a longer annotation it is able to keep it in a single piece. We also note that unlike Table 4, the minimum of 0 for SVM^{hmm} here corresponds to actual performance and not a failure case.

In Table 6, we show a broken-down example of multiple predicted annotations according to CRFsuite (PA), with respect to a gold annotation, illustrating how each sentence was identified by the algorithm. What we see is several of the incorrect labels are potentially reasonable outcomes. What we would not expect is 'and' and '57' to be of particular importance and are likely the result of text spanning multiple pages with OCR errors and possible poor sentence segmentation. On the other hand, there are several, much longer, sentences that we do not see a reason why the model decided to not predict a relevant label for them. Accordingly, while in practice, the former case is probably admissible, the latter is most certainly not.

On average a gold standard annotation spans 5.2 sentences which when combined with Table 6 would lead us to believe that a non-trivial proportion of CRFsuite's and SVM^{hmm}'s performance in Table 5 results from poor sentence segmentation and/or OCR errors. The best remediation for those errors is to improve OCR and sentence segmentation, which is outside the scope of this work. That being said, both approaches are also incorrectly labelling longer, relevant sentences as well.

In summary, the results from this section and the previous one indicate that while approaches similar to VW may be tempting and do look good at first glance, digging into the failure cases reveals a number of deficiencies. Solutions to such deficiencies would also be applicable to non-VW approaches (e.g., if two relevant sentences have less than X non-relevant ones in between them, treat the entire set of sentences as relevant). Consequently, we find that we have affirmed our choice of using CRFs and sequence-based representations of sentences as a viable and effective solution to the due diligence problem. Investigating further the utility of approaches like SVM^{hmm} may bear useful fruit but we have not seen anything to cause us to switch course from CRFs and CRFsuite.

7 LIMITATIONS

7.1 Hyper-Parameter Tuning

One of the most obvious flaws present in this work is that we have used largely untuned systems. Our tuned CRFsuite parameters do stem from tuning on initial proof of concept data but may themselves be out of date for many of our more recent topics (including those in this work). Accordingly, while the results in this work might be practically "good enough," they are almost certainly not the best possible results. Determining the optimal configurations for these systems would likely yield an improved user experience for anyone attempting to address due diligence problems.

Similarly, our features were also tested in a similarly ad hoc manner to get a minimum viable product working reasonably well. This is not to say that we do not think we have arbitrarily dismissed any particular set of features, only that we have not been as rigorous formally testing all possible combinations as we might otherwise have been.

Optimizing hyper-parameters and feature sets does have one particular caveat in our own particular use case, the resulting algorithm must still be reasonably quick. Increasing featurization time, training time, and memory used throughout are all practical constraints that we have had and continue to have to deal with. Given the relatively slim margins to improve effectiveness in some cases, we would be hesitant to adopt any approach that would substantially increase the time users have to wait to get results out of the system.

As a case in point, if we consider the topic that SVM^{hmm} failed to predict any relevant labels on and solely cut the 'e' parameter in half (i.e., to 0.5) then we approximately triple the training time from 39 minutes to 113 minutes. Though the resulting model produces an annotation-level recall of 0.69 and a precision of 0.97, this barely beats CRFsuite (PA) in precision (0.96) and is much worse on recall (0.81). Subsequently, it is unclear if increases of that magnitude are worth it, especially if we have to fine tune for every new topic we encounter. Such fine-tuning may be possible for technical experiments but is not something we believe a user of a due diligence system (i.e., a lawyer) ought to have to perform or be willing to wait for if they have preconceived notions of how long it ought to take.

P	Sentence Text
-1	21.9 Information: miscellaneous
1	The Parent shall supply to the Agent (in sufficient copies for al the Lenders, if the Agent so requests, or in electronic form if the Borrower so elects),
1	(a) at the same time as they are dispatched, copies of all documents dispatched by the Parent or any Obligors to its creditors generally (or any class of them);
-1	(b) promptly upon becoming aware of them, the material details of any litigation, arbitration, or administrative proceedings which are current, threatened in writing or pending against any member of the Group, and which, if adversely determined, are reasonably likely to have a Material Adverse Effect or which would involve an uninsured liability, or a potential or aleged uninsured liability, exceeding US$10,000,000 (or its equivalent in other currencies);
-1	(c) promptly, such information as the Security Agent may reasonably require about the Charged Property and compliance of the Obligors with the terms of any Transaction Security Documents;
-1	and
-1	57
1	(d) promptly on request, such further information regarding the financial condition, assets and operations of the Group and/or any member of the Group (including any requested amplification or explanation of any item in the financial statements, budgets or other material provided by any Obligor under this Agreement and/or details of any changes to the Senior Management of the Parent or the Borrower as any Finance Party through the Agent may reasonably request.

Table 6: (P)redicted labels (1: relevant, -1: not relevant) and the associated OCR'd sentence text for a gold annotation that had multiple predicted overlapping annotations according to CRFsuite (PA). Several predicted labels appear to be resulting from poor sentence segmentation and possibly OCR errors, while others are outright failures.

7.2 Beyond Passages

This work has primarily focused on identifying sentences that are pertinent to a set of particular information needs for the purpose of expediting the due diligence process, however, sentences may be too coarse. Indeed, things like "purchase price," or "start date," or the parties involved in a legal document are very fine-grained details that annotating entire sentences may not capture well. Adapting the methods presented herein for that level of granularity is an important next step in aiding the due diligence process. Such details may be able to be pulled out with more traditional CRF models (e.g., sequences of tokens) or we may be able to apply normalization and heuristics on top of sentence labelling.

On the other hand, when we consider form-based documents (e.g., more finance-oriented), it may not be the case that sentence-based methods will not work due to a lack of sentences. While we might like a "one model type fits all" approach, it is unlikely to be the case.

7.3 Document Diversity

As was discussed earlier, most documents[4] in the Kira Systems collection focus on American and Canadian law and are primarily credit agreements. Extending this dataset to account for the other jurisdictions' document types is crucial and one we plan to undertake. By extending the dataset in this way, there is also an interesting question as to whether or not you can extend existing models to new jurisdictions simply by adding additional training data or whether a completely new model would work best.

Along with new jurisdictions comes a bigger problem, different languages. The documents in our dataset are primarily in English and so this limits exploration of how different languages will affect performance. But the question still remains, do these techniques extend to other languages? Are we limited to Latin-based alphabets? What about Arabic, Chinese kanji, or Cyrillic? How do we handle documents with mixed-languages? These are all open questions worthy of exploration.

[4]A smattering of documents originate in the UK.

8 FUTURE WORK

Due to a lack of space and our goal of presenting a competitive baseline, we have elided discussion of deep learning and neural network methodologies. We do acknowledge that Recurrent Neural Networks and Long-Short-Term Memory networks have been applied to other sequence tagging tasks [16, 19, 27] with success, and so is an area we are actively exploring.

There are however, practical considerations to deep networks that have caused us to focus on these core methods. Namely, amount of training data and training time. In much of the published literature, training high quality neural networks can require massive amounts of data and potentially weeks or months of training time, even with GPUs. As we have seen, we can currently do well with relatively few example documents, and would need deep methods to do well under these same restrictions. More importantly, users can sometimes become frustrated, even now, when it takes an hour or two to train a large model. We would be hesitant to tell them that this would become even longer using neural networks for potentially minimal gains.

We might see this type of information extraction being used to extract useful information from research publications. An example of this might be determining how many papers from the last three decades have used, not just mentioned, a particular dataset.[5] Being able to collect and analyze this type of information may allow meta-studies to determine impact of core ideas that might otherwise have been overlooked.

Finally, it is still an open question on how to quantify risk in the identified clauses. In particular, do these models have to be bespoke? While there may be some general themes in what makes a clause risky, one client may be more or less averse to particular types of risk than another. Furthermore, once we can quantify risk, we might then wonder how to present this information to the user. Do we show the riskiest topics with the riskiest passages first or do we show the riskiest documents first? Is such document-level risk measured cumulatively or averaged across topics? There remains

[5]Example inspired from Susan Dumais' keynote [10] at the 25th Anniversary of TREC.

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.*