A Reliable and Accurate Multiple Choice Question Answering System for Due Diligence

Radha Chitta and Alexander K. Hudek
Kira Systems
Toronto, Canada
{radha.chitta,alex}@kirasystems.com

ABSTRACT

The problem of answering multiple choice questions, based on the content of documents has been studied extensively in the machine learning literature. We pose the due diligence problem, where lawyers study legal contracts and assess the risk in potential mergers and acquisitions, as a multiple choice question answering problem, based on the text of the contract. Existing frameworks for question answering are not suitable for this task, due to the inherent scarcity and imbalance in the legal contract data available for training. We propose a question answering system which first identifies the excerpt in the contract which potentially contains the answer to a given question, and then builds a multi-class classifier to choose the answer to the question, based on the content of this excerpt. Unlike existing question answering systems, the proposed system explicitly handles the imbalance in the data, by generating synthetic instances of the minority answer categories, using the Synthetic Minority Oversampling Technique. This ensures that the number of instances in all the classes are roughly equal to each other, thus leading to more accurate and reliable classification. We demonstrate that the proposed question answering system outperforms the existing systems with minimal amount of training data.

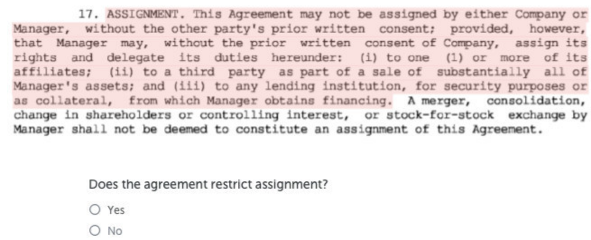
CCS CONCEPTS

• **Information systems → Question answering**; Document filtering; *Clustering and classification.*

KEYWORDS

Question Answering, Imbalance Handling, Due Diligence

ACM Reference Format:
Radha Chitta and Alexander K. Hudek. 2019. A Reliable and Accurate Multiple Choice Question Answering System for Due Diligence. In *Seventeenth International Conference on Artificial Intelligence and Law (ICAIL '19), June 17–21, 2019, Montreal, QC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3322640.3326711



17. ASSIGNMENT. This Agreement may not be assigned by either Company or Manager, without the other party's prior written consent; provided, however, that Manager may, without the prior written consent of Company, assign its rights and delegate its duties hereunder: (i) to one (1) or more of its affiliates; (ii) to a third party as part of a sale of substantially all of Manager's assets; and (iii) to any lending institution, for security purposes or as collateral, from which Manager obtains financing. A merger, consolidation, change in shareholders or controlling interest, or stock-for-stock exchange by Manager shall not be deemed to constitute an assignment of this Agreement.

Does the agreement restrict assignment?

○ Yes
○ No

Figure 1: A sample question *Does the agreement restrict assignment?* with the evidence highlighted in the contract and a yes/no answer.

1 INTRODUCTION

The objective of due diligence in mergers and acquisitions (M&A) law [18] is to examine legal documents and assess the risk in the potential mergers and acquisitions. We formulate the problem of extracting this information from the contract as a multiple-choice question answering problem, and present a question answering system which can predict answers to user-defined questions after identifying the text in the contract most relevant to the question. For instance, an important factor lawyers are interested in is if the contract can be assigned to a third party without the consent of all the original parties to the contract. This can be posed as the question *"Does the agreement restrict assignment?"* with a yes/no answer (Refer Figure 1). The proposed question answering system is trained to first identify relevant passages from the contract which potentially contain the answer to this question, and then learn a classification model to predict the answer based on the content of these passages. Most existing systems employed for multiple choice question answering based on the content of documents suffer from the drawback of imbalanced training data, i.e. it is not possible to find nearly equal number of instances of each type of the answer for a given question to perform the training. For instance, there may be more examples of contracts where the answer to the question *"Does the agreement restrict assignment?"* is *yes* when compared to *no*. Existing systems cannot handle such imbalance and often require tens of thousands of contracts to accurately learn to predict the answers for multiple choice questions.

In the proposed system, we explicitly handle the imbalance in the training data by generating synthetic examples of the minority categories, using the **S**ynthetic **M**inority **O**versampling **TE**chnique (SMOTE) [3, 20]. This technique, frequently used for imbalance handling in classification, generates new examples as linear combinations of instances of the minority classes, until the number of instances in the majority and minority classes are nearly equal to each

ICAIL '19, June 17–21, 2019, Montreal, QC, Canada

ACM ISBN 978-1-4503-6754-7/19/06…$15.00
https://doi.org/10.1145/3322640.3326711

other. The incorporation of this technique in the proposed question answering system ensures that the training can be performed accurately with minimal number of examples. Our preliminary results demonstrate the accuracy and efficiency of the proposed system in answering multiple choice questions.

The remainder of the paper is organized as follows: In Section 2, we outline some of the existing methods for question answering and review the related literature. In Section 3, we describe the proposed question answering system, and present our preliminary results and conclusions in Sections 4 and 5 respectively.

2 PRIOR WORK

There are two primary types of methods employed for question answering: (i) using a structured knowledge base manually built by domain experts, and (ii) using natural language processing techniques. In the former approach to legal question answering, lawyers spend hours building data structures and rule based systems for answering questions. These systems are neither cost-effective nor robust enough to handle different types of questions.

Natural language processing techniques, on the other hand, analyze the text of the documents and automatically learn to predict the answers for questions. They are more robust and scalable to large number of documents, thus reducing the time and the financial cost of the system. The proposed system falls into the latter category.

Most prior work based on natural processing techniques involves identifying parts of the documents which potentially contain the answer to the question and building a multi-class classifier over the (document, question, answer) triplets to find the answer to questions on previously unseen documents [6, 14]. Words or phrases in the documents and questions are represented as feature vectors and the parts of the document most relevant to the question are identified based on the similarity between the document sentences and the question. Lately, more complex embedding methods like deep neural networks are being employed to featurize the documents and questions, and for building classification models [5, 9, 11, 12, 19].

In [6], a ranking Support Vector Machine [8] is employed to rank the paragraphs in a set of legal articles against the question, in the order of their relevance. A convolutional neural network classifier is then trained to determine the answer based on the most relevant excerpts. The question and the articles are converted to TFIDF (**T**erm **F**requency **I**nverse **D**ocument **F**requency) vectors [17] before being input to the ranking SVM. In [9], the sentences in the articles and the question are embedded using the *word2vec* [13] technique to get the feature vectors. These word2vec features, combined with additional linguistic features, are then input to a convolutional neural network classifier to obtain the answers. Both these works assume that the data set is balanced.

Attention-based recurrent neural networks [5, 11, 12, 19] first embed the sequence of words in a sentence, using a learned embedding matrix, to obtain the feature vectors for each word. The embedding matrix may or may not be the same for the question and the document sentences. The word sequence within the sentence affects each word's vector representation due to its dependence on the previous words. In [5] and [11], the feature vector for each word is additionally multiplied by learned weights to signify the importance of the word in the sentence. The feature vectors thus

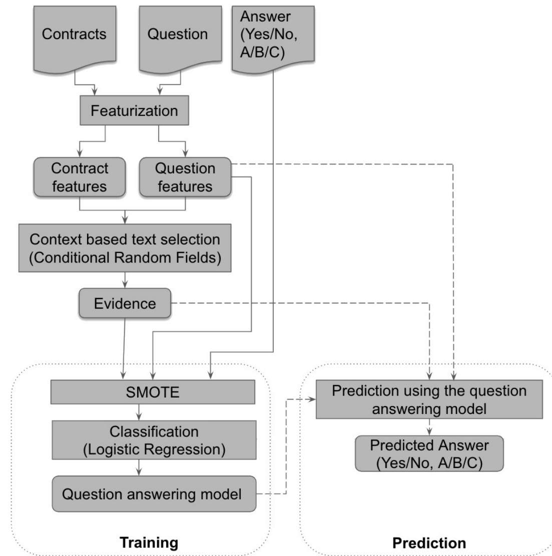


Figure 2: Training and prediction procedures for question answering.

obtained are then encoded using a recurrent neural network layer containing either Gated Recurrent Units [4] or Long Short-Term Memory Units [7]. The encoded representation of the question is matched with those of the document sentences to find the sentences most relevant to the question, and then passed through another recurrent neural network layer to obtain the answer. The advantage of these deep neural network architectures is that they can be employed to answer factoid based questions (with free text as answers) in addition to multiple choice questions. However, they suffer from the inherent drawback of all deep neural networks, i.e. they require massively large amount of training data to learn the question answering model accurately.

In summary, the existing systems for question answering require large amounts of training data, and are not sufficiently accurate and reliable because they are not designed to handle imbalanced data. The proposed question answering system explicitly incorporates measures to handle the data imbalance leading to more accurate and reliable results with minimal amount of training data.

3 METHODOLOGY

The proposed question answering system takes in as input a set of contracts, and a set of multiple choice questions to be answered based on the text of the contracts. For training the system, it also requires the true answers for the questions. Figure 2 describes the training and prediction procedures used by the proposed system.

3.1 Training

The training process (Refer Figure 2) starts with the training data containing a set of (*contracts*, *questions*, and *answers*) triples. The answers may be either yes/no, or have multiple options such as A/B/C, etc. The contracts and questions are first featurized, as described below, to obtain the vector representations for each of them.

The *evidence* is the piece of text in the contract which contains the answer to the question. This evidence can either be marked manually in the contract by the lawyer, or extracted automatically. In our implementation, we use a context based text selection algorithm called **C**onditional **R**andom **F**ields (CRF) [10]. The evidence, question and the answers together form the training data for the next phase of learning. This training data is passed through the SMOTE algorithm and then the logistic regression algorithm to obtain the question-answering *model*. Each of these steps are described in detail below:

- **Featurization:** Each sentence in the contract is treated as a set of tokens (words or phrases or n-grams), and converted into a unique d-dimensional TFIDF vector representation x reflecting the frequency of each word/phrase. The TFIDF of a token t in sentence s, given the set of all sentences S in the input data, is defined by

$$tfidf(t,s,S) = tf(t,s) * idf(t,S), \tag{1}$$

where

$$idf(t,S) = \log\left[\frac{(1+n)}{(1+df(t,S))}\right] + 1, \tag{2}$$

$tf(t,s)$ is the frequency of term t in sentence s,
$df(t,S)$ is the number of sentences that contain the term t, and n is the total number of sentences in S.
The questions are similarly converted to unique vector representations. Each of the answers is represented by a class label y, where $y \in \{l_1, l2, \ldots, l_K\}$, the K possible answer categories.

- **Evidence Extraction:** After the contracts and questions are featurized, these features are passed through a context based text selection algorithm, which extract the portions of the contract which are most relevant to the question. We employ the technique proposed in [16], which uses the CRF algorithm for evidence extraction, and has been shown to yield the best performance for this task.
A CRF is a probabilistic generative model for assigning labels to sequential data. Given a sequence of data points $X = \{x_1, x_2, x_3, \ldots\}$ and their labels $\{z_1, z_2, z_3, \ldots\}$, a CRF models the joint probability of the data and the labels, given by

$$P(z, X, \lambda) = \frac{\exp\sum_i\sum_j \lambda_j F_i(X, i, z_{i-1}, z_i)}{\sum_z \exp\sum_i\sum_j \lambda_j F_i(X, i, z_{i-1}, z_i)}, \tag{3}$$

where $F_i(X, i, z_{i-1}, z_i)$ is a feature function dependent on the current position of the token and labels of the current and previous data points, and $\{\lambda_j\}$ are a set of learnt weights associated with the feature functions. The parameters are learnt by maximizing (3).
In our scenario of evidence extraction, the data points x_i represent the feature vectors of the sentences, and the labels z_i represent whether the sentence is relevant to the question ($z_i = 1$) or not relevant ($z_i = 0$).
In an alternative implementation of the proposed question answering system, the evidence could be marked manually by a lawyer. This evidence is then featurized to obtain its vector representation. However, this method of implementing the system is only feasible on a small scale, and not recommended.

- **Synthetic Minority Oversampling TEchnique (SMOTE):** The differentiating factor in the proposed question answering system is that it produces more accurate and reliable prediction models than the existing systems with relatively small amount of training data. This is due to the incorporation of the SMOTE algorithm [3, 20] for improving the quality of the training data. A data set is imbalanced if the number of instances of each category of data are not approximately equal to each other. For instance, for a *"Does the agreement restrict assignment?"* question with a yes/no answer, if there are 90 instances where the answer is yes, and just 10 instances where the answer is no, this data set is said to be imbalanced. This lack of balance in the training data set leads to inaccurate and unreliable models. A large amount of training data, ranging in tens of thousands of examples, would be required to counter the effects of such data imbalance.
To limit the amount of training data required, while also countering the effect of data imbalance, we incorporate the SMOTE synthetic data generation algorithm. The SMOTE algorithm generates synthetic instances as linear combinations of existing minority class instances, as follows:
  - For each minority instance x, find k of its nearest neighbors belonging to the minority class. In our implementation, we employ the $k-d$ tree based k-nearest neighbor algorithm [1] to find the nearest neighbors.
  - Randomly select a neighbor x_{nn} from among the k-nearest neighbors, and generate a new instance x_{new} as

$$x_{new} = x + r\,(x_{nn} - x), \tag{4}$$

  where r is a uniform random number between 0 and 1.

These two steps are repeated until the number of minority class instances are approximately equal to the number of majority class instances. The class label y_{new} for each of the synthetic instances x_{new} is set equal to the class label of x. Unlike other methods for imbalance handling, SMOTE generates synthetic instances which are more representative of the minority class, and has been shown to yield the best results.

- **Multi-class classification:** The new data set obtained after SMOTE is passed through the logistic regression classification algorithm [2]. Logistic regression is a probabilistic classifier which maximizes the conditional probability that given a data point x_i, its label y_i equals l_k. This conditional probability is given by

$$P(y_i = l_k|x_i) = \frac{\exp f_k(x_i)}{\sum_k \exp f_k(x_i)}, \tag{5}$$

where $f_k(x_i) = w_k x_i + b_k$, w_k is a d-dimensional weight vector for class l_k, and b_k is the corresponding bias term. The parameters w_k and b_k are learnt by maximizing (5).

Question	No. of *yes* instances	No. of *no* instances
Q1: Is notice required to assign	10	119
Q2: Does assignment require consent	75	40
Q3: Does the agreement restrict assignment	96	19
Q4: Does the agreement renew automatically	72	128

Table 1: Yes/No Questions in the Experimental Data set

3.2 Prediction

Figure 2 describes the procedure of predicting an answer for a question based on the content of the contract. The contract and question are featurized, and the evidence is extracted using the CRF, as described in Section 3.1. The feature vector for the evidence x is then used to calculate the probability that the label for the data point x is l_k, for each of the K labels, using (5). The label y^* is then obtained as

$$y^* = max_k P(y = l_k|x). \quad (6)$$

Finally, the answer corresponding to y^* is displayed to the user.

4 EXPERIMENTAL RESULTS

In this section, we evaluate the efficacy of the proposed multiple choice question answering system, and compare it to the existing techniques for multiple choice question answering.

4.1 Data Description

Our experimental data set consists of 147 legal agreements of different types including service agreements, intellectual property agreements, supply agreements, etc. The agreements were obtained from the publicly available data sources EDGAR (Electronic Gathering, Analysis, and Retrieval system) and SEDAR (System for Electronic Document Analysis and Retrieval). These sources contain material documents filed by publicly traded companies in the United States and Canada, respectively.

We examine the performance of the proposed system on four yes/no questions[1] listed in Table 1. Clearly, this data set is highly imbalanced and small in size[2]. The evidence for each of the questions is obtained by converting the sentences in the agreements and the question to bi-gram TFIDF vectors, and applying the CRF algorithm on these vectors. We used the CRFsuite software [15] for the CRF algorithm implementation. Our in-house annotators (consisting of law students, contract lawyers, and in-house senior lawyers) then tagged each question and evidence pair with the actual yes/no answer. This formed the ground truth for our classifier training and evaluation.

4.2 Evaluation Strategy

We performed a random $80 - 20$ split on the instances of each of the four questions to form the training and test data sets, ensuring that there were both yes and no instances in both the splits. We trained the proposed system using the training set and evaluated its performance on the test set.

[1]Note that although we use simple yes/no questions to demonstrate its performance, the proposed system can also be applied to questions with multiple choices as answers.
[2]Sophisticated deep neural network architectures such as attention based recurrent neural networks cannot be trained accurately on data sets of such small size.

Question			Q1	Q2	Q3	Q4
QA systems without data imbalance handling	Train	A	90.4	83.7	83.5	99.5
		P	86.5	83.7	83.5	100.0
		R	100.0	100.0	100.0	98.6
		F1	92.7	91.1	91.0	99.3
	Test	A	57.6	82.6	83.5	94.0
		P	57.5	82.6	83.5	96.1
		R	93.5	100.0	100.0	84.7
		F1	70.2	90.4	90.9	89.8
Proposed QA system	Train	A	97.5	99.8	100.0	99.4
		P	98.4	100.0	100.0	98.2
		R	97.3	99.7	100.0	98.2
		F1	97.8	99.9	100.0	99.1
	Test	A	70.4	92.2	85.2	96.0
		P	71.4	100.0	100.0	92.5
		R	82.1	90.4	82.1	90.0
		F1	76.2	94.9	90.1	95.0

Table 2: Comparison of the accuracy (A), precision (P), recall (R) and F1 scores (in percentage) of the proposed question answering system with and without the incorporation of SMOTE, on the training and test data sets. Unlike the proposed question answering system, the existing question answering systems do not incorporate any data imbalance handling techniques. The proposed QA system uses SMOTE to balance the data, due to which its performance is significantly improved.

(a) QA systems without data imbalance handling

	Predicted: Yes	Predicted: No
Actual: Yes	19	0
Actual: No	4	0

(b) Proposed QA system

	Predicted: Yes	Predicted: No
Actual: Yes	18	1
Actual: No	0	4

Table 3: Test confusion matrices for the question *Does assignment require consent?*

The objective of our work is to show that by using SMOTE for data imbalance handling, the performance of the question answering system is improved. Therefore, we compare the performance of the system with and without the incorporation of SMOTE to balance the data set, and evaluate the improvement in performance. Note that the proposed system without the incorporation of SMOTE is equivalent to any of the existing multiple choice question answering systems, which first learn to extract the sentences relevant to the question from the contract and then directly learn a classifier on this imbalanced data. In order to prove that SMOTE synthesises instances representative of the minority class, we also balanced the data set using randomly generated instances and then learnt the classifier. We call this method the random balance method.

We set the parameter k (number of neighbors for the SMOTE algorithm) to 6. We evaluate the performance of the systems in terms of the following four measures :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (7)$$

$$Precision = \frac{TP}{TP + FP}, \quad (8)$$

$$Recall = \frac{TP}{TP + FN}, \quad (9)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}, \quad (10)$$

where the terms TP, TN, FP, and FN are defined as

- TP (True Positive): Number of actual *yes* answers predicted as *yes*.
- TN (True Negative): Number of actual *no* answers predicted as *no*.
- FP (False Positive): Number of actual *no* answers predicted as *yes*.
- FN (False Negative): Number of actual *yes* answers predicted as *no*.

4.3 Results

Table 2 records the performance of the proposed system with and without the incorporation of SMOTE for balancing the data. The first eight rows show the training and test evaluation measures for the imbalanced data (equivalent to the performance of any of the existing question answering systems) and the last eight rows show the training and test measures of the proposed system, which incorporates SMOTE.

The proposed system is clearly better performing in terms of the accuracy and F1 measures, proving that incorporation of SMOTE for data balancing helps immensely. This is especially true for the question *Is notice required to assign?*, where the imbalance in the majority and minority instances is the largest (12 : 1 ratio). The test accuracy for this question increases by about 13%, due to the incorporation of SMOTE for imbalance handling. When the data is imbalanced, the logistic regression classifier overfits the data, leading to the low accuracy on the test data. The random balance method also results in degraded performance. For instance, the test accuracy for the question *Is notice required to assign?* is only 64% for this method. The results of this method for all the questions are not included due to lack of space.

We observe that the test recall suffers moderately on the first three questions in the proposed system. This is because the false negative rates are slightly higher in the proposed system, as seen in the confusion matrix for the question *Does assignment require consent?* in Table 3. However, the increase in the precision (increase in true positive and reduction in false positive rates) more than compensates for this reduction in recall.

In summary, we observe that the proposed system significantly improves the performance of multiple choice question answering, as a result of explicitly handling the data imbalance issue. It is very useful, especially in scenarios where instances of some answers are hard to come by, which is often the case in the legal domain.

5 CONCLUSIONS

Though question answering based on the content of documents has been extensively studied in the fields of artificial intelligence and machine learning, there have been few works and products which apply the same to the legal domain. The primary reason that the sophisticated machine learning techniques for question answering cannot be directly applied to legal documents (contracts) is that the contracts are diverse, leading to insufficient examples for training and data imbalance.

In this article, we proposed a multiple choice question answering system which *explicitly* handles the imbalance in the data by generating synthetic examples. The system assesses the legal document, identifies excerpts within the document which contain the answer to a given question and then chooses the appropriate answer from among the available options. We demonstrate the effectiveness of the proposed system on simple yes/no questions, which are useful for performing due diligence in mergers and acquisitions.

In the future, we plan to extend the proposed question answering system to answer multi-answer questions (questions with two or more options as an answer) and factoid based questions (questions with free text as answers).

REFERENCES

[1] Jon Louis Bentley. 1975. Multidimensional Binary Search Trees Used for Associative Searching. *Commun. ACM* 18, 9 (1975), 509–517.
[2] Christopher M Bishop. 2008. *Pattern Recognition and Machine Learning*. Springer.
[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.
[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).
[5] Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit, Illia Polosukhin, Alexandre Lacoste, and Jonathan Berant. 2017. Coarse-to-fine question answering for long documents. In *Proceedings of the Association for Computational Linguistics*, Vol. 1. 209–220.
[6] Phong-Khac Do, Huy-Tien Nguyen, Chien-Xuan Tran, Minh-Tien Nguyen, and Minh-Le Nguyen. 2017. Legal question answering using ranking SVM and deep convolutional neural network. *arXiv preprint arXiv:1703.05320* (2017).
[7] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.
[8] Thorsten Joachims. 2006. Training linear SVMs in linear time. In *Proceedings of the International Conference on Knowledge Discovery and Data mining*. 217–226.
[9] Mi-Young Kim, Ying Xu, and Randy Goebel. 2015. Applying a convolutional neural network to legal question answering. In *JSAI International Symposium on Artificial Intelligence*. 282–294.
[10] John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning*. 282–289.
[11] Huayu Li, Martin Renqiang Min, Yong Ge, and Asim Kadav. 2017. A Context-aware Attention Network for Interactive Question Answering. In *Proceedings of the International Conference on Knowledge Discovery and Data mining*. 927–935.
[12] Fenglong Ma, Radha Chitta, Saurabh Kataria, Jing Zhou, Palghat Ramesh, Tong Sun, and Jing Gao. 2017. Long-Term Memory Networks for Question Answering. *arXiv preprint arXiv:1707.01961* (2017).
[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. 3111–3119.
[14] Amit Mishra and Sanjay Kumar Jain. 2016. A survey on question answering systems with classification. *Journal of King Saud University Computer and Information Sciences* 28, 3 (2016), 345–361.
[15] Naoaki Okazaki. 2007. CRFsuite: a fast implementation of Conditional Random Fields. http://www.chokkan.org/software/crfsuite.
[16] Adam Roegiest, Alexander K Hudek, and Anne McNulty. 2018. A Dataset and an Examination of Identifying Passages for Due Diligence. In *International ACM SIGIR Conference on Research & Development in Information Retrieval*. 465–474.
[17] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press.
[18] 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. *Richmond Journal of Law & Technology* 21, 2 (2015), 5.
[19] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems*. 2440–2448.
[20] Boyu Wang and Joelle Pineau. 2016. Online bagging and boosting for imbalanced data streams. *IEEE Transactions on Knowledge & Data Engineering* 1 (2016).