Passing a USA National Bar Exam: a First Corpus for Experimentation

Biralatei Fawei, Adam Wyner, and Jeff Pan
Department of Computing Science, University of Aberdeen, Aberdeen, United Kingdom
biralatei.fawei,azwyner,jeff.z.pan@abdn.ac.uk

Abstract

Bar exams provide a key watershed by which legal professionals demonstrate their knowledge of the law and its application. Passing the bar entitles one to practice the law in a given jurisdiction. The bar provides an excellent benchmark for the performance of legal information systems since passing the bar would arguably signal that the system has acquired key aspects of legal reason on a par with a human lawyer. The paper provides a corpus and experimental results with material derived from a real bar exam, treating the problem as a form of textual entailment from the question to an answer. The providers of the bar exam material set the Gold Standard, which is the answer key. The experiments carried out using the 'out of the box' the Excitement Open Platform for textual entailment. The results and evaluation show that the tool can identify wrong answers (non-entailment) with a high F1 score, but it performs poorly in identifying the correct answer (entailment). The results provide a baseline performance measure against which to evaluate future improvements. The reasons for the poor performance are examined, and proposals are made to augment the tool in the future. The corpus facilitates experimentation by other researchers.

Keywords: textual entailment, bar exam, legal reasoning, natural language processing

1. Introduction

Bar exams, which are extensive in-depth examinations about legal information and reasoning, provide a key watershed by which legal professionals demonstrate their knowledge of the law and its application. Passing the bar entitles one to practice the law in a given jurisdiction and topic, and it validates the examinee's knowledge of the law. We can say, then, that the bar exam encapsulates a range of legal knowledge. Thus, the bar provides an excellent benchmark for the performance of legal information systems which attempt to represent and reason with the law, since passing the bar would arguably signal that the system has acquired key aspects of legal reason on a par with a human lawyer.

This paper presents a dataset and a first experiment with bar exam material derived from the United States Multistate Bar Examination (MBE) material, provided by the National Conference of Bar Examiners (NCBE) [1]. The Gold Standard (the correct answers) is provided by the NCBE. The paper reports a textual entailment study on this US bar exam material, running the Excitement Open Platform (EOP)[2] for textual entailment 'out of the box' (Dagan et al., 2009). In the experiment, we treat the the relationship between the question and the multiple-choice answers as a form of textual entailment. The results and evaluation show that the tool can identify wrong answers (non-entailment) with a high F1 score, though it performs poorly in identifying the correct answer (entailment). The results provide a baseline performance measure against which to evaluate future improvements. The reasons for the poor performance are examined, and proposals are made to augment the tool in the future.

The study is cast in a more general context of question answering for the law. Question answering is an automatic way of determining the right answer as a response to a question presented in natural language form (Harabagiu and Moldovan, 2003). Among the many varieties of questions, we treat bar exam questions as a form of 'Yes/No' question; given background information and a statement about that information, is the statement true or false with respect to the background information? Question answering is useful in the legal domain, which faces the challenge of finding and determining the right statement given some background information; this is particularly daunting given the volume and complexity of legal information. The problematics only increase with legal reasoning from resources found on the Internet. Broadly, it is pressing to find approaches to extract and process information so as to identify the correct answer to a given question.

The research adopts the textual entailment technique to determine if a given text *t*, known as the theory, *entails* another text *h*, known as the hypothesis (Dagan et al., 2009). The concept of entailment used in this technique is broader than logical concept of *logical entailment* - given t would one accept (or reject) h. Gold Standard corpora are devised to provide data for experiments. For our purposes, the bar exam question constitutes the theory and the question to be picked as the correct answer constitutes the hypothesis; the answers provided are the Gold Standard. The entailment classification is based on semantic relatedness and mutual consistency. The essence is to find out semantic relatedness and mutual consistency between a legal text in natural language form as question and some answer in natural language form, where semantic relatedness and mutual consistency bear on the terminology of the texts.

In this research, we attempt to ascertain the semantic relatedness and mutual consistency between pairs of text question and answer from a large legal corpora. The findings are, in brief, that the textual entailment tool that is used is largely successful identifying answers that are *not* entailed by the question, but largely unsuccessful in identifying answers that *are*. The findings provide a baseline for future work, which would augment the 'out of the box' system with specifically legal information. In previous work

[1] http://www.ncbex.org/
[2] http://excitement-project.eu/

(Fawei et al., 2015), we have presented a precursory corpus along with preliminary results of the application of EOP to that corpus. The novel contributions of this paper are the descripton and presentation of a new, larger, and 'cleaner' corpus for legal text NLP along with a more refined result from EOP. The significance of the work is that by laying a baseline, it provides a means to measure future incremental improvements to a successful legal question answering system. Such a system would, in our view, have broad and deep implications for the access to and practice of the law.
The rest of the paper is organized as follows. Section 2. describes the legal corpus and its features. Section 3. explains the textual entailment tool, the Excitement Open Platform (EOP), as well as some selected associated algorithms. Section 4. presents the experiment and the results. Section 5. discusses some related works in the domain, while Section 6. wraps up the research discussion with observations and future work.

2. Corpus Description

The *National Conference of Bar Examiners* in the United States prepares and administers the Multistate Bar Exam (MBE) every year to thousands of aspiring lawyers throughout the country. The MBE is an obligatory, 6-hour, 200-question multiple-choice test given in every US state but Louisiana. It accounts for 40-50% of an aspiring lawyers bar exam score (other exams taking up the other 50-60%). In 2014, 73,088 examinees took the test; the mean scaled score (out of 200) was 140.4; approximately 29% of examinees did not pass the exam.[3] The exam questions (in the most recent exam) cover the legal spectrum: Constitutional Law, Contracts, Criminal Law and Procedure, Evidence, Real Property, Torts, and Civil Procedure. Thus, the MBE is a broad and deep exploration of the examinee's knowledge of the law as it applies across the US.
A legal corpus was gathered from NCBE materials and prepared for a textual entailment exercise on the Excitement Open Platform. The original dataset contains one hundred questions, each with four possible answers out of which the candidate must pick the correct one; the NCBE provided an answer key to the materials. Given some modifications discussed below, the original dataset was developed into pairs of theories and hypothesis, where each question was paired with one possible answer, yielding four hundred theory-hypothesis pairs.[4] The correct (Gold Standard) answer is indicated as *entailment* and the wrong answer as *nonentailment*. Analysed this way, there is a bias of nonentailment-entailment in the ratio 3:1. The Gold Standard corpus contains 66306 words with 3071 sentences of which the sentences that are theory contains 2671 sentences with minimum of 4 and maximum of 13 sentences while the sentences that are hypothesis is 400 sentences.
An example original question with answers a.-d. is:

> After being fired from his job, Mel drank almost a quart of vodka and decided to ride the bus home. While on the bus, he saw a briefcase he mistakenly thought was his own, and began struggling with the passenger carrying the briefcase. Mel knocked the passenger to the floor, took the briefcase, and fled. Mel was arrested and charged with robbery.
>
> Mel should be
>
> a. acquitted, because he used no threats and was intoxicated.
> b. acquitted, because his mistake negated the required specific intent.
> c. convicted, because his intoxication was voluntary.
> d. convicted, because mistake is no defense to robbery.

For the textual entailment task, the material must be presented in a particular XML format, where the theory appears as a whole, then the hypothesis in a single sentence. Constructing the material for EOP processing required significant manual preprocessing. As illustration, we mention issues with this example.
First, in the original format, we have one question followed by four possible answers, whereas in the XML format, each question can be followed by only one answer (*Question division*). Second, the original background portion includes part, e.g. "Mel should be", of what conceptually ought to be part of the hypothesis (what is entailed), along with the main verb (*full proposition*). Third, the possible answers portion includes part, e.g. a justification "because he used no threats and was intoxicated", of what conceptually ought to be part of the theory (*justification*). Fourth, the justification must, when moved to the theory, maintain a reasonable narrative "flow" (*narrative*). Fifth, the original is not in XML (*XML*). Finally, due consideration must be given to the derived format so that it is meaning preserving (*meaning preserving*); in our examples, meaning preservation requires that we identify the core inference and put all "background" information into the theory.
Given these considerations, we have samples of derived questions:

> <pair id="7A" entailment="NONENTAILMENT" task="QA">
>
> <t>After being fired from his job, Mel drank almost a quart of vodka and decided to ride the bus home. While on the bus, he saw a briefcase he mistakenly thought was his own, and began struggling with the passenger carrying the briefcase. Mel knocked the passenger to the floor, took the briefcase, and fled. Mel used no threats and was intoxicated. Mel was arrested and charged with robbery.</t>
>
> <h>Mel should be acquitted.</h>
>
> </pair>

[3] Data accessed September 05, 2015 from `http://www.kaptest.com/bar-exam/courses/mbe/multistate-bar-exam-mbe-change`.
[4] The analysed dataset is available upon request

<pair id="7B" entailment="ENTAILMENT" task="QA">

<t>After being fired from his job, Mel drank almost a quart of vodka and decided to ride the bus home. While on the bus, he saw a briefcase he mistakenly thought was his own, and began struggling with the passenger carrying the briefcase. Mel knocked the passenger to the floor, took the briefcase, and fled. Mel was arrested and charged with robbery. Mel's mistake negated the required specific intent.</t>

<h>Mel should be acquitted.</h>

</pair>

A range of other structural issues of the text were identified and controlled for in order to produce a corpus that conceptually matches the original:

- Meta comments about the exam question, e.g. "Assume that...." and "Which of the following is correct?"
- References to other cases, e.g. "As applied in Long's case....".
- Pronominal anaphora, e.g. "It is a generally applicable statute....", where the modification might disrupt the anaphoric chain.
- Changes in verbal form, e.g. "...as applied..." becomes a main verb "is applied".
- Scope of particles, e.g. "if any" must be attached to relevant elements.
- "Yes" and "No" in original questions to refer to positive and negative forms of the hypothesis.
- Subordinate clauses in h are made into main clauses in t, e.g. "In a suit for conversion by Homeowner against Neighbor...." to "Homeowner makes a suit against Neighbor for conversion."

3. Excitement Open Platform EOP Description

In this section, we briefly outline the Excitement Open Platform (EOP). The EOP is an open source platform made available for both scientific and technological community for textual inference. The essence of the platform is to deliver an automatic means of identifying textual entailment between a pair of texts. The EOP platform was developed and implemented to provide a common framework for users and developers to experiment with textual analysis using multilingual resources and a variety of algorithms (Magnini et al., 2014).

The EOP currently contains five different entailment decision algorithms: BIUTEE, Edit Distance, Textual Inference Engine, P1EDA and AdArte. We experimented with Edit Distance and the Textual Inference Engine.

3.1. Edit Distance

The ED algorithm uses a series of mapping operations in order to map the entire semantic content of the theory to the hypothesis in order to determine entailment. The mapping operations are edit operations such as *delete*, *insert* and *substitution*. Each of these operations are associated with a cost value, in which the probability of entailment between text pairs could be derived by taking an inverse proportion of the edit distance between the text pairs (Padó et al., 2014; Magnini et al., 2014). The cost of each operation is given as 0 for match, 0 for delete, 1 for insert, and 1 for substitution. The algorithm measures semantic similarities between pairs of texts by measuring token edit distance and tree edit distance. It applies some similarity measures such as Word overlap, Cosine similarity, and Longest common sequence to measure similarity between theory and hypothesis. An entailment decision is taken based on the number of operations that led to making the theory and hypothesis to be identical, concentrating its findings based on the minimal number operations that lead to the goal state. The edit distance algorithm uses a threshold of approximately 0.5741 and accuracy measure of 0.6575 to determine entailment between text pairs.

3.2. Textual Inference Engine TIE

The TIE algorithm is similar to that of the edit distance algorithm, but in addition checks entailment based on relatedness/similarity and mutual consistency, determining whether there is an inherent directionality between the given theory and hypothesis. A confidence value is assigned. It uses analysis on bag-of-words along with syntactic and semantic dependency information. Relatedness is a measure of similarity/difference of concepts, sentences and words measure (McInnes and Pedersen, 2013). For our purposes, relatedness measures the extent to which a pair of sentences are related to each other. The similarity measure quantifies similarities between two concepts based on the information they contain (Pedersen et al., 2004), which is obtainable with the help of the WordNet lexical database. The bag-of-words takes the theory and hypothesis pair of the corpus as a set of words and returns a score based on similarity and relatedness from the pair. This measurement technique relies on VerbOcean for extraction of related verbs as well as WordNet for expansion of related words and Google Normalized Distance (GND) computation of distance with respect to terms (Padó et al., 2014). The bag-of-word feature returns two scores which are within 0 and 1 for both theory and hypothesis. The syntactic information compares the theory and hypothesis pair based on dependency parse trees. Bags of dependency triples are extracted from the text pairs and computed for normalised values for the theory and hypothesis. The normalised values lie between 0 and 1, which are used for the identification of relatedness between the text pairs. If the normalised value is 0 then there is no relatedness, but if it is 1 then identical. This feature models word dependency in a sentence. The knowledge resources used in this component are VerbOcean, WordNet and GND and operated on MSTParser.

4. Experiments and Results

We applied the EOP to our corpus of MBE questions. A number of trials were carried out to ascertain the degree, measured by standard Accuracy (A), Precision (P), Recall (R) and F1 measures, to which the various algorithms (Edit Distance and Text Inference Engine) could be used to reliably identify theories with entailed (E) from nonentailed (NE) hypotheses. Using the TIE algorithm on the corpus of 400 pairs, out of the 100 Gold Standard entailment examples, the system was able to confirm 23 actual entailments while failing to accurately identify the remaining 77 (see Tables 1-2); it would may be that the ratio of entailment to non-entailment sentences biased the algorithm. Out of the 300 Gold Standard nonentailment examples, the system incorrectly identified 69 as entailments while confirming 231 as nonentailment. The TIE algorithm had the highest entailment result with 0.903026 confidence value and a highest nonentailment result with 0.501849 confidence value. The Edit Distance algorithm (ED) performed worse on the corpus of 400 pairs, correctly identifying only 11 out of the 100 entailment examples and 22 out of the 300 nonentailment examples (see Tables 3-4). The algorithm had the highest entailment result with 0.574176 confidence value and a highest nonentailment result with 0.002747 confidence value.

In order to avoid bias, the dataset was redistributed with each correct pair along with one wrong pair, constituting three different datasets each with two hundred pairs. The algorithms were reapplied on the redistributed dataset. The results were the same or slighly worse in comparison with the initial dataset of 400 pairs; to conserve on space, we have suppressed these results.

To summarise, the algorithms used in the EOP have not succeeded in coming close to getting enough the correct answers to pass the USA national bar exam. However, they can reliably identify the wrong answers.

	E	**NE**
E	23	77
NE	69	231

Table 1: Contingency Table for TIE algorithm

	A	**P**	**R**	**F1**
NE	63.5	75.0	77.0	75.987
E	63.5	25.0	23.0	23.958

Table 2: Results from TIE algorithm

	E	**NE**
E	11	89
NE	22	278

Table 3: Contingency Table for ED Result

	A	**P**	**R**	**F1**
NE	72.25	75.749	92.667	83.358
E	72.25	33.333	11.0	16.541

Table 4: Results from EDA

5. Related Work

The most closely related work is (Kim et al., 2013; Tran et al., 2013).[5] In (Tran et al., 2013), an analysis is applied to 51 legal questions on the Japanese National Pension Law; the focus is on retrieval of relevant texts rather than textual entailment per se. The approach seems to be that closely related legal information ought to have closely related references to other texts, which are retrieved and used to augment the content of the texts being examined. Textual similarity and logical structure analysis are used to determine the relationship between question and answer. They report an improved performance over approaches without retrivial, with an accuracy of about 60%. The sample of data is relatively small (51 questions); the role of the augmented texts and logical structure analysis are difficult to gauge. Finally, the underlying analysis is done on Japanese and not on Bar Exam questions, so the comparison to US Bar Exam questions is indirect. More directly relevant is (Kim et al., 2013), which works with a corpus of Japanese/Korean Bar exam questions, which include legal articles and questions. Questions are analysed in terms of negations and complexity. A rule-based system for Japanese legal reasoning is applied with results of about 60% accuracy for all questions. The structure of the material (language, question and articles) is different from the US Bar Exam; the tool is highly specific; moreover, the relationship between the source natural language text and the rule-based analysis is unclear.

Question-answering and textual inference have long been studied, though not with application to legal corpora. For question-answering, inference has been used (Lin and Pantel, 2001; Segura-Olivares et al., 2013), though noisy situations reduces performance. An answer validation technique that utilizes the subsequence kernel method has been implement for machine learning for question answering (Wang and Neumann, 2008). A linear-chain Conditional Random Field (CRF) has been integrated into Tree Edit Distance for extracting answers (Yao et al., 2013). A lexical and syntactic feature similarities technique for determining textual entailment between a pair of texts has been applied (Pakray et al., 2011). A tree kernel approach is used to drive a greedy search routine to decide textual entailment between a pair of texts (Heilman and Smith, 2010). A similarity metrics measure was adopted (Rios and Gelbukh, 2012) in recognizing textual entailment. The research adopted string based metrics, chunking and named entities recognition as well as shallow semantic metrics for recognizing textual entailment. In (Bobrow et al., 2007), a rule-based approach is described to determine entailment and contradiction between a pair of texts. A semantic inference mechanism alongside cost-based approximation for deciding en-

[5] http://webdocs.cs.ualberta.ca/~miyoung2/jurisin_task/index.html

tailment between a pair of texts is presented in (Bar-Haim et al., 2007). The framework operates on parse trees to generate new trees based on entailment rules to decide if the hypothesis is entailed in the text. While these approaches require further improvement, it would be worth exploring in the EOP context whether they would augment the results when applied to legal texts.

6. Discussion

The paper reports the development of a corpus and the application of the EOP to determine textual entailment relations between questions and answers on US legal bar exams. The results show some success in identifying entailment and nonentailment pairs of sentences. From the experiments, it is clear that while recognition of nonentailment is rather high, the recognition of entailment is poor.

One of the key observations to emerge from this study is the importance of logical reasoning in making entailment determinations. Using bags of words based on enrichment of lexical information or syntactic dependencies is not sufficient. Consider the following two examples (with simplified questions):

> Question 1: ....Tina decided that the house needed improvement, and she paid cash to have installed standard- sized combination screen/storm windows, a freestanding refrigerator to fit a kitchen alcove built for that purpose, a built-in electric stove and oven to fit a kitchen counter opening left for that purpose, and carpeting to cover the plywood living room floor....
>
> A. The court should decide that Tina may remove none of the items.
>
> B. The court should decide that Tina may remove only the refrigerator.
>
> C. The court should decide that Tina may remove all items except the carpet.
>
> D. The court should decide that Tina may remove all of the items.
>
> Question 2: ....Proposal A would eliminate the insanity defense altogether. Proposal B would retain the defense but place on the defendant the burden of proving insanity by a preponderance of the evidence. Opponents of the reforms argue that the proposals would be unconstitutional under the ....
>
> A. proposals would be unconstitutional.
>
> B. Neither proposal would be unconstitutional.
>
> C. Proposal A only would be unconstitutional.
>
> D. Proposal B only would be unconstitutional.

In these examples, the algorithms determine that all four of the possible answers are entailed by the question. The reason is that all the possible answers are closely semantically related to the text. The algorithms only use explicit textual information or augmentations provided by the resources. Several other examples in our data set fall under this sort of problemantic.

Another issue identified in course of the experiment is that the materials used to augment the textual information, e.g. VerbOcean and WordNet, lack the sorts of legal legal information and reasoning that is required. For example, the following possible answers not only refer to a relevant legal document, but also to some reasoning extracted from it:

> ....the original jurisdiction of the Supreme Court as defined by Article III ....
>
> ....appellate jurisdiction of the Supreme Court, because Article III states....
>
> ....in support of the EPAs request is that Article III precludes....
>
> ...support of the EPAs request is that Article III provides that....

To decide entailment in this case requires constitutional knowledge. With the current application, nonentailment is fairly reliably identified since this relies on the textual *difference* between theory and hypothesis, whereas for entailment, examples textual similarity is not reliable as the theory and hypothesis can be rather distinct, yet semantically related. In future work, we will develop legal resources that can serve to augment textual entailment tools so as to improve the results of a textual entailment tool.

The work reported here is novel in that it is a first, open, well-developed corpus of legal textual on US Bar Exams which is specifically designed to address matters of inference. The results lay a baseline for future developments.

7. Acknowledgments

The authors appreciate the permission granted by the Multistate Bar Examination organisation to work with their bar exam materials. The first author gratefully acknowledged support by Niger Delta University through the Tertiary Education Trust Fund (TETFund).

8. Bibliographical References

Bar-Haim, R., Dagan, I., Greental, I., Szpektor, I., and Friedman, M. (2007). Semantic inference at the lexical-syntactic level for textual entailment recognition. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 131–136. ACL.

Bobrow, D. G., Condoravdi, C., Crouch, R., de Paiva, V., Karttunen, L., King, T. H., Nairn, R., Price, L., and Zaenen, A. (2007). Precision-focused textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 16–21. ACL.

Dagan, I., Dolan, B., Magnini, B., and Roth, D. (2009). *Natural Language Engineering*, 15:i–xvii, 10.

Fawei, B., Wyner, A., and Pan, J. (2015). Passing a USA national bar exam - a first experiment. In *Legal Knowledge and Information Systems - JURIX 2015: The Twenty-Eighth Annual Conference, Braga, Portual, December 10-11, 2015*, pages 179–180.

Harabagiu, S. and Moldovan, D. (2003). Question answering. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, pages 560–582. Oxford University Press.

Heilman, M. and Smith, N. (2010). Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. In *HLT-NAACL 2010*, pages 1011–1019. ACL.

Kim, M.-Y., Xu, Y., Goebel, R., and Satoh, K. (2013). Answering yes/no questions in legal bar exams. In *New Frontiers in Artificial Intelligence - JSAI-isAI 2013 Workshops, Japan, October 2013*, pages 199–213.

Lin, D. and Pantel, P. (2001). Discovery of inference rules for question-answering. *Natural Language Engineering*, 7(04):343–360.

Magnini, B., Zanoli, R., Dagan, I., Eichler, K., Neumann, G., Noh, T.-G., Pado, S., Stern, A., and Levy, O. (2014). The excitement open platform for textual inferences. In *Proceeedings of 52nd Annual Meeting of the ACL*, pages 43–48. ACL.

McInnes, B. and Pedersen, T. (2013). Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text. *Journal of biomedical informatics*, 46(6):1116–1124.

Padó, S., Noh, T.-G., Stern, A., Wang, R., and Zanoli, R. (2014). Design and realization of a modular architecture for textual entailment. *Journal of Natural Language Engineering*, 21:167–200, 3.

Pakray, P., Bandyopadhyay, S., and Gelbukh, A. (2011). Textual entailment using lexical and syntactic similarity. *International Journal of Artificial Intelligence and Applications*, 2(1):43–58.

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at HLT-NAACL 2004*, pages 38–41. ACL.

Rios, M. and Gelbukh, A. (2012). Recognizing textual entailment with similarity metrics. *Research in Computing Science*, 58:337–347.

Segura-Olivares, A., García, A., and Calvo, H. (2013). Feature analysis for paraphrase recognition and textual entailment. *Research in Computing Science*, 70:119–144.

Tran, O. T., Ngo, B. X., Nguyen, M., and Shimazu, A. (2013). Answering legal questions by mining reference information. In *New Frontiers in Artificial Intelligence - JSAI-isAI 2013 Workshops, Japan, October 2013*, pages 214–229.

Wang, R. and Neumann, G. (2008). Using recognizing textual entailment as a core engine for answer validation. In *Advances in Multilingual and Multimodal Information Retrieval*, pages 387–390. Springer.

Yao, X., Durme, B. V., Callison-burch, C., and Clark, P. (2013). Answer extraction as sequence tagging with tree edit distance. In *HLT-NAACL*, pages 858–867.