Noise over Fear of Missing Out

Legal professionals prefer recall over precision for AI-assisted legal information extraction tasks

Johannes Schleith
Johannes.Schleith@tr.com
Thomson Reuters Labs
London, UK

Nina Hristozova
Nina.Hristozova@tr.com
Thomson Reuters Labs
Zug, Switzerland

Brian Cechmanek
Brian.Cechmanek@tr.com
Thomson Reuters Labs
London, UK

Carolyn Bussey
Carolyn.Bussey@tr.com
Thomson Reuters
London, UK

Leszek Michalak
Leszek.Michalak@tr.com
Thomson Reuters Labs
Zug, Switzerland

ABSTRACT

Natural language processing (NLP) techniques for information extraction commonly face the challenge to extract either 'too much' or 'too little' information from text.

Extracting 'too much' means that a lot of the relevant information is captured, but also a lot of irrelevant information or 'Noise' is extracted. This usually results in high 'Recall', but lower 'Precision'. Extracting 'too little' means that all of the information that is extracted is relevant, but not everything that is relevant is extracted – it is 'missing' information. This usually results in high 'Precision' and lower 'Recall'.

In this paper we present an approach combining quantitative and qualitative measures in order to evaluate the end-users' experience with information extraction systems in addition to standard statistical metrics and interpret a preference for the above challenge. The method is applied in a case study of legal document review. Results from the case study suggest that legal professionals prefer seeing 'too much' over 'too little' when working on an AI-assisted legal document review tasks. Discussion of these results position the involvement of User Experience (UX) as a fundamental ingredient to NLP system design and evaluation.

CCS CONCEPTS

• Human-centered computing → HCI design and evaluation methods; **• Computing methodologies → Information extraction**.

KEYWORDS

User centric evaluation of AI-based systems, Precision, Recall, Task Support

1 INTRODUCTION

Traditionally Artificial Intelligence (AI) assistance of information extraction tasks has been seen mainly as a technological challenge [1][2]. In this paper we argue that adoption of systems for such tasks also requires careful consideration of the end-user experience, as well as the evaluation of 'Trust', the need for explanation as well as optimizing for Precision or Recall depending on the use case and desired experience. Precision is the amount of true positives divided by the sum of all true positives and false positives identified by the model. Likewise, Recall is the fraction of all true positives to the sum of all true positives and false negatives identified by the model [3] [4].

Research in Human Computer Interaction (HCI) has developed methodology to evaluate the end-user experience with intelligent systems, such as Usability (i.e. effectivity, efficiency, satisfaction)[5], Task Completion and Trust, according to IS 9241-210:2019 [6] and ISO 9241-11:2020 [7][8] respectively.

2 METHOD

Our experiment aims to assess the perceived quality of an information extraction system, investigate perceptions of Trust and a notion of 'Completeness' and ultimately understand end-user preference for optimization for Recall or Precision in AI-assisted legal information extraction tasks.

2.1 System

An experimental system automatically reviews legal documents and extracts legal language that could potentially be relevant for legal professionals to answer 'Specific Questions' about such documents. This study focuses on the evaluation of the results rather than the exact workings of the technology. In a nutshell, the system used for this experiment applies information extraction techniques based on NLP models that had been trained with carefully selected and annotated training data similar to other research [9][10].

In our study we used a model that had been trained on real estate leases and Specific Questions such as *"Who is the landlord?"* or *"What are the repair obligations?"*. As common for such systems, the results vary in levels of extraction quality, due to the amount and quality of training data per specific question, the specific documents under review as well as the required level of quality of any given legal document review.

Mensch und Computer 2021, Workshopband, Workshop on User-Centered Artificial Intelligence (UCAI '21)

https://doi.org/10.18420/muc2021-mci-ws02-290

2.2 Annotation Task

In order to assess the system, we recruited 20 participants (8 practicing lawyers, 12 legal editors), with legal training and relevant work experience, from an internal pool of volunteers. Participants used the system to review a number of real estate leases in two rounds, 95 documents in total. The first round included 20 Specific Questions, the second round another 13 Specific Questions. The system provided a user interface (UI) to either accept or reject each automated extraction or add annotations per Specific Question.

2.3 Metrics

Comparing automated extractions against annotations by end-users and domain experts allowed us to calculate standard statistical metrics, such as Recall, Precision, and F1, per question, which are commonly applied to evaluate the quality of information extraction [11]. In addition, we captured perceived 'Task Support' by asking *"How helpful was the system for your review and answering the question?"* on a 5-point Likert scale (1=*"Very unhelpful"*, 5=*"Very helpful"*) per document and per Specific Question. Scores are reported as averaged Task Support per Specific Question. This approach allowed us to contrast statistical metrics with self-reported Task Support.

In addition, we followed up with 10 participants in semi-structured interviews about their impressions of the helpfulness of the tool.

Moderation involved questions such as *"What makes a good/bad question"*, *'How do you feel about seeing more results (with some relevant answers and some irrelevant noise)"*, *"How do you feel about seeing fewer results (while possibly missing out on relevant answers)"*.

2.4 Results

Comparing scores for Recall and Precision to Task Support suggest that participants evaluated Specific Questions, that showed many, possibly noisy results, as more helpful than Specific Questions that showed fewer, possibly incomplete results (see figure 1), for this AI-assisted legal information extraction task.

A multiple linear regression on Recall and Precision on Task Support found a significant equation ($F(2, 30) = 31.8, p << 0.01, R^2 = 0.67$). Recall significantly predicted Task Support ($B = 2.68, p << 0.01$), while Precision did not significantly predict Task Support ($B = -0.01, p = 0.97$). A significant interaction effect of Recall*Precision on Task Support was found ($B = 1.07, p << 0.01$).

This can be interpreted such that Specific Questions with high Recall were perceived as "more helpful" than questions with low Recall. While low or high Precision did not have such an impact.

Qualitative feedback from follow-up interviews show why Recall more strongly correlates with Task Support than Precision. Participants described a general 'Fear of Missing Out' and being afraid of missing something. In the legal domain there are high consequences for providing bad advice. In contracts even seemingly insignificant changes to text can have large monetary impact for a client. When extractions methods did not yield any relevant text, just showing 'no results' made participants anxious (6 of 10 participants). They suggested that more help in confirming a non-answer or finding the relevant information would improve their impression (5 of 10).

For some Specific questions, participants expected an answer to exist in the document so a non-answer was perceived as an obvious 'miss' by the system. In other Specific questions, non-answers were often considered risky and some wanted a way to confirm that a question was 'not addressed in the document' (5 of 10). A few mentioned that, seeing 'no answers' decreased their trust in the system over time (2 of 10).

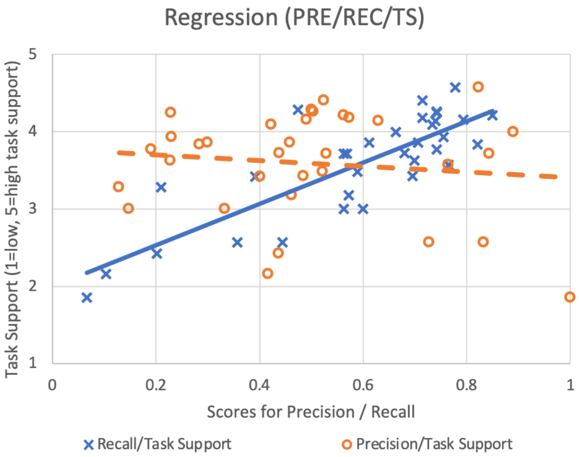

Figure 1: Regression analysis. Each dot represents a 'Specific Question' positioned on its score for Recall, Precision and Task Support

Participants desired a notion of Completeness of results and coverage of sources. Given the novelty of the extraction methods in the legal domain, they felt that they would need to manually double check everything (6 of 10). During user tests we observed participants often double-checked AI-assisted search with a simpler search like Ctrl-F (4 of 10) or by reading the whole document (4 of 10) to "make sure they cover it all".

3 DISCUSSION

In this study we show an approach that compares a user-centered evaluation of AI text extraction results with data-driven metrics. We further present significant results from the application of this method in a case study and interpret a preference for Recall over Precision for search and information extraction tasks in the legal domain. We argue that it is crucial to involve user-centered evaluation of AI text extraction output early in the process, in order to optimize data science methods towards validated user goals and preferences.

Future work should explore the evaluation of end-users' notion of Completeness and design approaches to communicate such Completeness in coverage of sources and identification of results.

Balancing Precision and Recall is a recurring challenge in various domains. More work might investigate preferences in other high stakes domains, such as medical, regulation or financial applications of AI. The field of HCI provides methodology for the evaluation of static content and taxonomies (e.g. Card Sort, Lostness etc.). Further research could explore a more robust framework and methods for the evaluation of information extraction and dynamically created content in a similar fashion.

REFERENCES

[1] Jakub Piskorski and Roman Yangarber. *Information Extraction: Past, Present and Future*, pages 23–49. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[2] Marie-Francine Moens, Caroline Uyttendaele, and Jos Dumortier. Intelligent information extraction from legal texts. *Information & Communications Technology Law*, 9(1):17–26, 2000.

[3] Nancy Chinchor and Ph D. Muc-4 evaluation metrics. In *In Proceedings of the Fourth Message Understanding Conference*, pages 22–29, 1992.

[4] Kai Ming Ting. *Precision and Recall*, pages 781–781. Springer US, Boston, MA, 2010.

[5] Maximilian Speicher. What is usability? a characterization based on iso 9241-11 and iso/iec 25010, 2015.

[6] Ergonomics of human-system interaction — Part 210. Standard, International Organization for Standardization, Geneva, CH, July 2019.

[7] Ergonomics of human-system interaction — Part 110. Standard, International Organization for Standardization, Geneva, CH, March 2020.

[8] Nigel Bevan, Jim Carter, Jonathan Earthy, Thomas Geis, and Susan Harker. New iso standards for usability, usability reports and usability measures. In Masaaki Kurosu, editor, *Human-Computer Interaction. Theory, Design, Development and Practice*, pages 268–278, Cham, 2016. Springer International Publishing.

[9] Michael J. Bommarito II, Daniel Martin Katz, and Eric M. Detterman. *LexNLP: Natural language processing and information extraction for legal and regulatory texts*, pages 216–227. Edward Elgar Publishing, Cheltenham, UK, 2021.

[10] mi-young Kim, Randy Goebel, Yoshinobu Kano, and Ken Satoh. Coliee-2016: Evaluation of the competition on legal information extraction and entailment. 11 2016.

[11] Monika Arora, Uma Kanjilal, and Dinesh Varshney. Evaluation of information retrieval: precision and recall. *International Journal of Indian Culture and Business Management*, 12(2):224–236, 2016.