Science and Technology Facilities Council
Hartree Centre

Collaborative Conveyancing
making complex simple

hartree.stfc.ac.uk

A Hybrid Approach for Information Extraction and Expert Action Recommendation using Fine-Tuned Base Models, Large Language Models and Knowledge Graphs

Robert Firth*[1], Jonny Palmer*[2], and Ruby George[1]
[1]the Science and Technology Facilities Council (STFC) Hartree Centre; [2]Collaborative Conveyancing Limited

Motivation

Conveyancing, *noun:* the legal process of transferring the ownership of real estate from one party to another.

Conveyancing involves a series of tasks and procedures that ensure the legal and financial aspects of a property transfer are properly executed. At the core of this are enquiries (“**Enquiries**”); requests for information or action for a relevant party.

Enquiries involve the exchange of **un-** or **semi-structured**, often complex questions and answers between conveyancers and form the basis of legal due diligence on the sale property, supporting the principle of ‘*caveat emptor*’ (buyer beware). Typical examples include requesting evidence of planning permission for extensions, copies of certificates for installations such as boilers, windows, etc and property title documents.

It is common to raise and resolve Enquiries via **email** and **attachment**, meaning that the caseload quickly becomes burdensome. The lack of a **common** platform, format, and the unstructured nature of the enquiries makes this a challenging problem for Natural Language Processing and Understanding (**NLP/NLU**). While contemporary Large Language Models (**LLMs**) excel in this domain, the regulated nature of the industry, and sensitivity of the process means that using unsupported LLMs carries an unacceptable **hallucination** risk.

We present a two-stage pipeline, an example of which is shown in **Figure 1**; first a **document parsing** and **classification** component “Enquiry Extraction” using rules-based segmentation and classification fine-tuned Transformer based on **DistilBERT**[1] (Cased); second, a “Task Identification” component that takes a reconstituted enquiry input and uses a fine-tuned **GPT3.5-turbo**[2] model to identify the action intents within the enquiry input and return a structured output. These tasks have attributes, such as Source Document and Subject that are then used to query the knowledge base for any relevant information relating to that task. This information is used to suggest candidate responses which can then be checked and emailed to the selected recipient

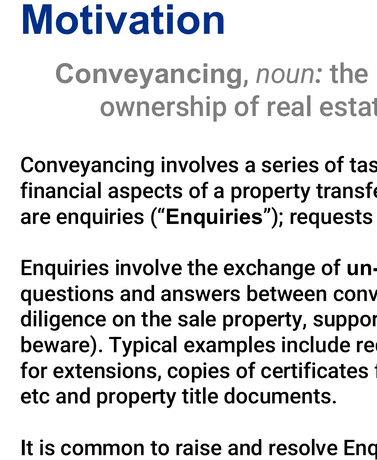

Figure 2 – *Left*: the Receiver-Operator Characteristic (ROC) curve of our fine-tuned classifier, showing the optimal decision threshold value of 0.28. *Right*: a confusion matrix showing the evaluation results of the same classifier

Acknowledgement: This research was funded by Innovate UK (Project 10054819) and delivered by the Science and Technology Facilities Council (STFC) Hartree Centre and Collaborative Conveyancing Limited. STFC is part of UK Research and Innovation.

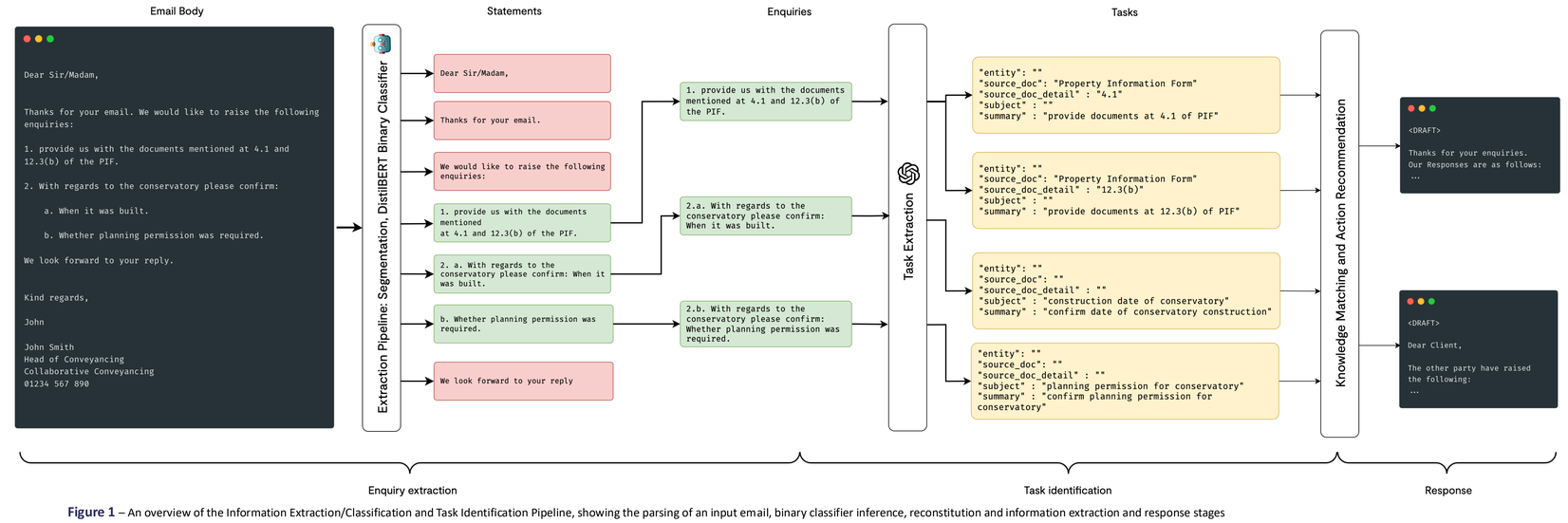

Figure 1 – An overview of the Information Extraction/Classification and Task Identification Pipeline, showing the parsing of an input email, binary classifier inference, reconstitution and information extraction and response stages

Data

Binary Classifier

A labelled dataset of **5776** enquiry annotations on **459** emails was generated by subject matter experts at Collaborative Conveyancing. This annotation was granular, with **1064** individual categories identified, so was down-sampled on a sentence-level to a **binary classification** task, resulting in a dataset with a 41:59 imbalance towards the negative (non-enquiry) class. See **Table 1**. The DistilBERT model was trained for 3 epochs, and achieved an Area Under Curve AUC=0.95; a **positive class recall** Recall=0.950 **and** Accuracy=0.888, **results shown in Figure 2.**

Split	N examples	Fraction
Train	12981	0.75
Test	2596	0.15
Eval	1731	0.1

Table 1 - Binary Classifier Training data

Knowledge Store

To ensure reliable Task Actions and Responses, subject matter experts assembled a knowledge bank, accessible in-memory in tabular or graph form. This consists of Source Documents, Aliases, special considerations, relevant information such as textbook responses or case law, definitions and considerations (**Table 2**).
This allows the retrieval of rich contexts to provide e.g. definitions or practical considerations for buyers and sellers, or for a model as part of a RAG system.

Entity Type	Count
Source Document	53
Source Document Details	215
Subjects	19
Responses	535
Considerations	166
Forward to Client	612

Table 2 – Knowledge Store

Instruction Pair Dataset

To extract the task information from the enquiries, **394 Expert-generated input-output pairs** were assembled to cover the most common enquiries, consisting of enquiry sequences and structured task output; extracting entities, source document, document identifiers, subject and summary to JSON. This was possible due to the formulaic language used by Conveyancers. Additionally **1000 synthetic input-output pairs** were generated using realistic language and slots for source documents, **30** examples per source. This set including examples where there is no intended action and both were used to fine-tune an instance of GPT3.5-turbo.

Output

Emails are split into statements which are scored by our **classifier**, with those classed as enquiries progressing through the pipeline, contextual information such as list signifiers are reconstructed before the Extracted Enquiries are passed to the fine-tuned GPT instance via a **prompt**, which generates the desired Task structure (**Figure 1**). This task structure is matched against the entities and validated responses using a cosine distance metric between that and the extracted subject and any source document. These are used to **recommend responses** to the end-user, such as **email drafts** to the client, or the sellers solicitor. The responses stored have been **human-curated** to ensure good coverage. If no match is found, no suggestions are made to avoid increasing the risk of damaging hallucinations.

Further Work

This system is currently being evaluated as part of **Collaborative Conveyancing**’s core product. As well as completing the end-to-end evaluation of the system further work will focus on flagging un-raised enquiries, building a graph representation of the conveyancing process, enriching retrieval with subject-object-predicate tags, emulation of the statement and enquiry stages, and using self-hosted language models.

References: 1. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv, abs/1910.01108*. 2. OpenAI (2024), ‘GPT-3.5 API’, API. Accessed: 2024-02-28. URL: https://openai.com/gpt-3 3. Honnibal, M., Montani, I., Van Landeghem, S. & Boyd, A. (2020), ‘spaCy: Industrial-strength Natural Language Processing in Python’. 4. Sadvilkar, N. & Neumann, M. (2020), PySBD: Pragmatic sentence boundary disambiguation, in ‘Proceedings of Second Workshop for NLP-OSS’, pp. 110–114.