Extracting Relevant Named Entities for Automated Expense Reimbursement

Guangyu Zhu
Institute for Advanced Computer Studies
University of Maryland, College Park
zhugy@umiacs.umd.edu

Timothy J. Bethea, Vikas Krishna
IBM Almaden Research Center
San Jose, CA 95120
{bethea, vikas}@us.ibm.com

ABSTRACT

Expense reimbursement is a time-consuming and labor-intensive process across organizations. In this paper, we present a prototype expense reimbursement system that dramatically reduces the elapsed time and costs involved, by eliminating paper from the process life cycle. Our complete solution involves (1) an electronic submission infrastructure that provides multi-channel image capture, secure transport and centralized storage of paper documents; (2) an unconstrained data mining approach to extracting *relevant* named entities from un-structured document images; (3) automation of auditing procedures that enables automatic expense validation with minimum human interaction.

Extracting *relevant* named entities robustly from document images with unconstrained layouts and diverse formatting is a fundamental technical challenge to image-based data mining, question answering, and other information retrieval tasks. In many applications that require such capability, applying traditional language modeling techniques to the stream of OCR text does not give satisfactory result due to the absence of linguistic context. We present an approach for extracting relevant named entities from document images by combining rich page layout features in the image space with language content in the OCR text using a discriminative conditional random field (CRF) framework. We integrate this named entity extraction engine into our expense reimbursement solution and evaluate the system performance on large collections of real-world receipt images provided by IBM World Wide Reimbursement Center.

Categories and Subject Descriptors: H.3.0 [Information Storage and Retrieval]: General; I.2.6 [Artificial Intelligence]: Learning; H.4.1 [Information Systems Applications]: Office Automation; I.7.5 [Document and Text Processing]: Document Capture – *Optical character recognition* (*OCR*);

General Terms: Algorithms, Experimentation, Design

Keywords: **N**amed entity extraction, learning, document layout analysis, conditional random fields.

KDD'07, August 12–15, 2007, San Jose, California, USA.

1. INTRODUCTION

Expense reimbursement is a tedious and laborious process for organizations of all sizes today. Even though policies and regulations defining the process vary across organizations and industries, corporate expense reimbursement is facing a set of common challenges. A complete solution to this problem requires technical innovations in the following three key areas.

(1) *A generalized paper-free framework for capturing, transporting, and storing paper documents in digital image form*

In spite of progress made with electronic tools, paper consumption in the office is growing and paper continues to inhibit business process innovation. Expense reporting is a classic example where web-based applications are mostly available in the organization, but the few remaining paper receipts continue to result in unnecessary costs and delays. The problem is that organizations require supporting receipt documents to prove the validity of submitted reimbursement claims. Currently, without any pervasive mechanism for electronic submission, paper receipts have to be mailed for centralized processing, together with printed cover sheets. To protect against the risk of loss in the mail, the package is often copied, which again creates more paper. Even worse than handling the expanding amount of paper, time to reimbursement remains at the speed of mailed package and manual processing rather than the speed of electronic transactions.

Fast technology shift in the printing industry from analog copiers to consolidated high-resolution digital multifunction devices (MFDs) enables us to close the paper–digital gap in the office environment using these pervasive image capturing and transporting channels. In our system, all it takes to submit paper receipts are a few easy steps: walking to the office MFD, authenticating yourself with your intranet password, selecting the appropriate menu option on the touch screen, and hitting the "big green button" to scan and submit receipts one by one.

(2) *Extraction of relevant named entities from receipt images with unconstrained layouts and formatting*

Steady progress in the optical character recognition (OCR) technology has significantly increased its usage for image-based information retrieval applications on digitized (scanned) documents [1, 2]. However, processing OCR text from un-structured contents like receipts poses serious challenges to traditional data mining approaches based on language modeling. First, text from receipts consists predominantly of terse streams of nouns. The lack of linguistic context, such as punctuation and language constructs, makes both syntactic and semantic analysis

very hard. Second, the output ASCII text from OCR does not contain useful page layout features, such as spatial block region and font information, which may be easily recognizable in the source image. Third, the OCR text quality is likely to be poor and full of errors, since receipts are typically printed by low-resolution impact printers (e.g. dot matrix printers) and are received in much worse physical condition compared to normal documents. These degradations on the source image are hard to recover and have significant impact on overall OCR performance. In practice, we encounter 6-10% character-level recognition errors on real-world receipt collections.

It is important to distinguish between the standard named entity extraction and recognition problem in the literature [3-5] and the one we are addressing here. In our task, the *query* to the *relevant* named entity in each category is equivalent to a *question*. Our objective is to find one *unique answer* that best answers the question using the presented context. For example, given a receipt document, we can ask what the name of the merchant is. If more than one merchant entities exist, the system needs to resolve such ambiguity and provide the one that is most likely to be the relevant answer using all available cues collectively.

Effective solution to both the entity extraction and question answering (QA) aspects of the problem requires integrating a multi-dimensional mixture of available features from language content and page layout, which is a research area with relatively little work in the literature. In addition, we prefer a formal model-based approach than a heuristic or rule-based approach. This paper focuses on the unconstrained image-based named entity extraction approach of our system.

(3) *Automation of auditing procedures that enables an organization to perform expense validation with minimum human interaction*

Many organizations are limited in their ability to audit expense reports, as it requires dedicated auditors to manually examine incoming receipts and judge their accuracy with information included in the associated report. This labor-intensive approach often causes an organization to downgrade their internal requirements for the percentage of submissions audited. For some organizations, especially those specializing in consulting or sales, which have a very high percentage of employees requiring travel, keeping costs down in expense processing generally requires a much lower rate of oversight than would be desired.

In this paper, we present an end-to-end expense reimbursement system that integrates an unconstrained data mining engine for extracting *relevant* named entities from un-structured document images. It dramatically reduces the elapsed time and costs involved in corporate expense reimbursement.

Our key data mining contributions include (1) a unified model based approach to jointly solving the named entity extraction and question answering aspects of the problem using conditional random fields (CRFs) [8]; (2) a formal framework for efficient probabilistic inference by learning contextual dependencies using both language and page layout features; and (3) demonstrating the feasibility and effectiveness of integrating rich feature sets available in the image space. The techniques presented in this paper can be generalized for solving broader image-based data mining problems, especially when OCR text reveals limited structural information.

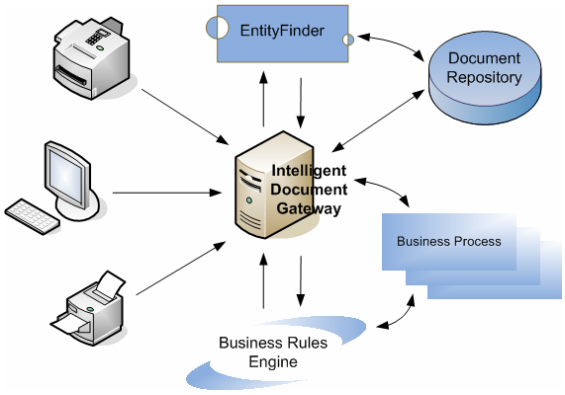

Figure 1: System architecture of our fully automated expense reimbursement system.

Using the CRFs relax the strong assumptions on the conditional independence of observations made in models like hidden Markov models (HMM), which can be unrealistic when dealing with overlapping features or long-term contextual dependencies in the observation sequence. Furthermore, conditional models like the CRFs offer a formal model-based framework of representing contextual dependencies, making efficient inference using dynamic programming, and training of parameters using convex optimization.

The remainder of this paper is structured as follows: In the next section, we describe the fully automated expense reimbursement system. We present an unconstrained approach to extracting relevant named entities from document images in Section 3. We discuss experimental results on collections of real-world receipts in Section 4 and conclude in Section 5.

2. SYSTEM OVERVIEW

In this section, we give a high-level overview of our expense reimbursement system and describe each technical component in its client-server architecture shown in Fig. 1. The client in this system can be any computer, multi-function device (MFD), fax machine, or other electronic device, which has built-in capability to transmit native image files. The application server running an IBM Intelligent Document Gateway (IDG) server directly interacts with the centralized document image repository, the named entity extraction engine (EntityFinder), the dynamic business rule engine server for automated auditing, and the whole range of associated business processes.

The electronic document submission, transport, and storage technologies presented here can serve effectively as the underlying infrastructure for a complex multi-document-type system, although our discussion focuses within the context of expense reimbursement in this paper. More elaborate description of the Intelligent Document Gateway (IDG) technology is presented in granted US patents [6, 7].

2.1 Electronic Submission

Our system provides multi-channel image capture, transport, and storage of paper documents. On the client side, users have several

options to submit paper receipts and scanned receipt images securely to the document gateway server:

1. **Multi-function devices (MFDs)**: The touch screens on the MFDs display customized UI interfaces for guiding users to scan and submit paper receipts, once they successfully authenticate by providing their corporate directory passwords.

2. **Web-based client**: The user can upload document image files directly from their computer to the server through a light-weight client. This web-based application allows the user to provide additional information associated with the submitted receipt document, including its language set and personal remainder, along with the image file.

3. **Desktop print-job**: This option allows the user to submit receipt document in native image format through their desktop printer queue.

The minimum meta-data transmitted along with a submitted image file includes the type of the document and the identifier that links the submitted receipt document to the corresponding expense claim. The receipt image and its associated meta-data are encrypted prior to transmission to the document gateway server via the corporate and pubic networks.

2.2 Named Entity Extraction

The capability to extract relevant named entities from document images is integrated into a data mining module called EntityFinder in our system, as shown in Fig. 2. At the lower level, EntityFinder handles images at their native formats (e.g. multi-page TIFF images) and provides support for higher-level functions, including document layout analysis and feature extraction, through interfaces with the OCR engine libraries. We present details of the unconstrained image-based data mining approach built into EntityFinder in Section 3.

2.3 Automated Auditing

Extracting relevant named entities from un-structured document images opens vast possibilities for business process automation. Our system makes use of a business rules engine to analyze the extracted data for relevancy within the context of automated expense auditing, and activate actions based on the result of the rule execution. The set of business rules are defined in XML and are dynamically configurable in the live system. The auditing actions include verification of extracted entities from receipt documents with reference to their corresponding entries in the expense reimbursement claim, flagging of instances of potential fraud, adherence to prescribed organizational policies such as meal limit, and calculation of more complex expense reimbursement activities, such as per diem verification or value-added-tax determination. Automation of these routine procedures enables a significantly higher rate of auditing and a much shorter turnaround time between submission and compensation, bringing tangible productivity gain and cost savings to the organization.

2.4 Document Archival

Once all business rules for automated auditing have been executed and resulting external business processes have been initiated, the set of extracted named entities, along with the source document, are stored in centralized repositories for archival. This

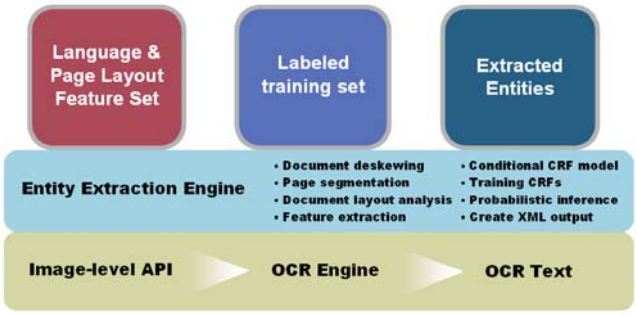

Figure 2: The named entity extraction module – EntityFinder.

task is required to conform to common business and legal requirements for document retention, for future data mining needs of the organization, and for auditing control purpose. The document archival process in our system is governed by a configuration file associated with the document process, which includes the type of repository adapter to use, link to the server and associated authentication data, and the descriptive information including table and column information for database access. This design allows multiple repositories at the backend as required by a given process, and can be flexibly adapted to organization-specific archival requirements.

3. EXTRACTION OF RELEVANT NAMED ENTITIES WITH CRFS

3.1 Related Work

Named entity recognition (NER) is an important task in deriving structured information from un-structured sources. Historically, it has been traditionally defined on un-structured text due to its root in Message Understanding Conferences [6]. NER capabilities have been demonstrated using un-structured text corpora from a wide range of domains, including identifying personal names and company names in newswire text [7], identifying titles and authors in on-line publications [8, 9], and identifying gene and protein names in biomedical publications [10, 11]. More recently, unsupervised NER results are reported on a massive corpus of domain-independent text from the web [12].

The vast majority of NER systems employ rule-based approaches and machine-learning-based approaches. Examples of rule-based systems in the literature include [13, 14]. Machine-learning-based approaches can be further divided into two main categories: classifier-based and Markov-model-based. Common choices of classifiers include decision trees, naïve Bayes, and Support Vector Machines (SVMs). In addition, studies in [15] and [16] have used classifier combination techniques in NER tasks. Markov-based models, including hidden Markov models (HMMs) [17], Maximum Entropy Markov Models (MEMMs) [18], and Conditional Random Fields (CRFs) [8], are well suited to problems that involve sequential analysis.

We choose CRFs as our framework as a result of the following motivations. First, CRFs relax the strict conditional independence assumptions of observations forced by generative models like HMMs for ensuring tractable inference [19]. This gives CRFs

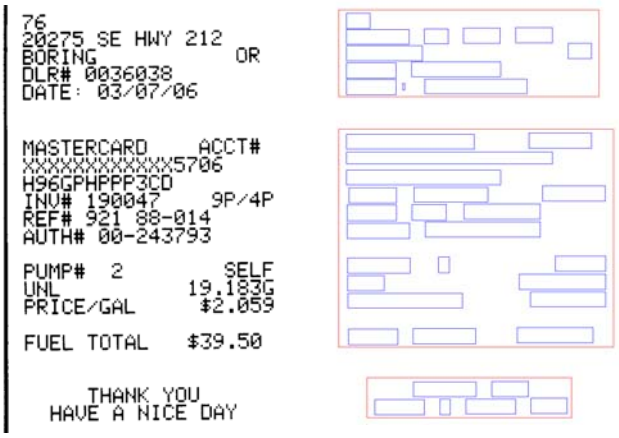

Figure 3: A receipt image (left) and the visualization of its segmentation results (right).

more flexibility to integrate complex, overlapping and non-independent feature sets that operate at multiple levels of granularity and multiple modalities. Second, modeling of conditional probabilities devotes model resources directly to the relevant task of label inference. It requires fewer labeled observation sequences, which generally leads to better generalization performance given limited training data. In addition, CRFs avoid the label bias problem exhibited by MEMMs and other discriminative Markov models based on directed graphical models [8]. A few studies have shown that CRFs outperform both MEMMs and HMMs on a number of real-world language related sequence labeling tasks [8, 20, 21, 22]. More recently, CRFs have also been extended to computer vision problems, including region classification [23] and human motion recognition [24].

3.2 Task and Approach

We consider the task of extracting the set of transaction-related NEs from receipt documents. Obviously, different levels of complexities are involved in extracting NEs of distinct natures. The limited variation in NEs like transaction amount, date, credit card number and merchant phone number can be handled effectively using regular expressions, in combination with rules. In this paper, we focus on the task of finding the set of NEs that have arbitrarily large variation (e.g. the merchant) and present a domain-independent approach to extracting such challenging NEs by effectively *exploiting context collectively* from the source image and its OCR text.

The application imposes three requirements:

1. Effectively handle document images with unconstrained layouts and formatting, since the system must be able to process all kinds of receipts.
2. Provide the most likely answer to each NE as inferred collectively from the context presented in the document.
3. Should not rely on large external dictionaries since it is not economic to create and maintain such dictionaries in practice. Furthermore, NEs on receipts commonly appear in various abbreviated forms that are hard to enumerate. In fact, even with this constraint lifted, the NE extraction task is not trivial, but presents a different set of problems. The challenges involved in improving NER performance using external dictionaries are discussed in [25].

The structural information derived collectively from page layout and language features is important to the NER task on un-structured documents like receipts, whose OCR text stream may not be sufficient. Fig. 3 shows a receipt example from a Union 76 gas station. The string "76" itself is most likely to be a number when it appears without context. However, people can find out that it refers to a merchant by examining the document layout and linguistic elements collectively.

A receipt document with arbitrary physical layout and formatting still conveys structural information in two aspects:

- Many logically and semantically related entities are geometrically placed within spatial proximities, even if the structure of them within the region is not obvious.
- The sequence of decomposed regions and the combination of layout and linguistic features within these regions reveals important contextual information collectively.

Our approach to extracting relevant NEs involves (1) decomposing the document image into homogeneous regions using page segmentation techniques; (2) learning the sequence of segmented regions and the containment of layout and linguistic features within regions by a discriminative conditional Markov model. CRFs provide a well-suited framework for integrating these non-independent features at multiple levels of granularity and multiple modalities.

Two page segmentation strategies can be employed to divide a general document image into homogeneous regions. One is to use a page segmentation algorithm. Representative page segmentation approaches from the document image analysis community include the Docstrum algorithm by O'Gorman [26] and the Voronoi diagram-based algorithm by Kise et al. [27]. Another approach is to use the OCR engine for this task. Fig. 4 shows the page segmentation results by the Docstrum algorithm[1] and the OCR engine, respectively. Each segmented region is plotted using a red bounding box. For better visualization of linguistic elements within each segmented region, we provide the OCR word segmentation result in the right sub-figure of Fig. 4 by drawing a blue bounding box around each segmented word.

Using the OCR engine directly to segment a document page has a few practical advantages as opposed to using a stand-alone algorithm. First, it makes region-level attributes easily accessible. Most leading commercial OCR packages offer region-level classification capabilities, which allow regions containing text, tables, graphics, and handwriting to be identified and processed accordingly. Second, it facilitates feature extraction from the segmented regions, including character-level attributes such as the coordinates of character borders on the image grid, font information, and recognition confidence. Last but not least, using

[1] Docstrum is a bottom-up page segmentation algorithm that is able to work on document images with non-Manhattan layout and arbitrary skew angles. It has limited capability to handle non-text regions and text zones with irregular font sizes and spacing, and tends to fragment them.

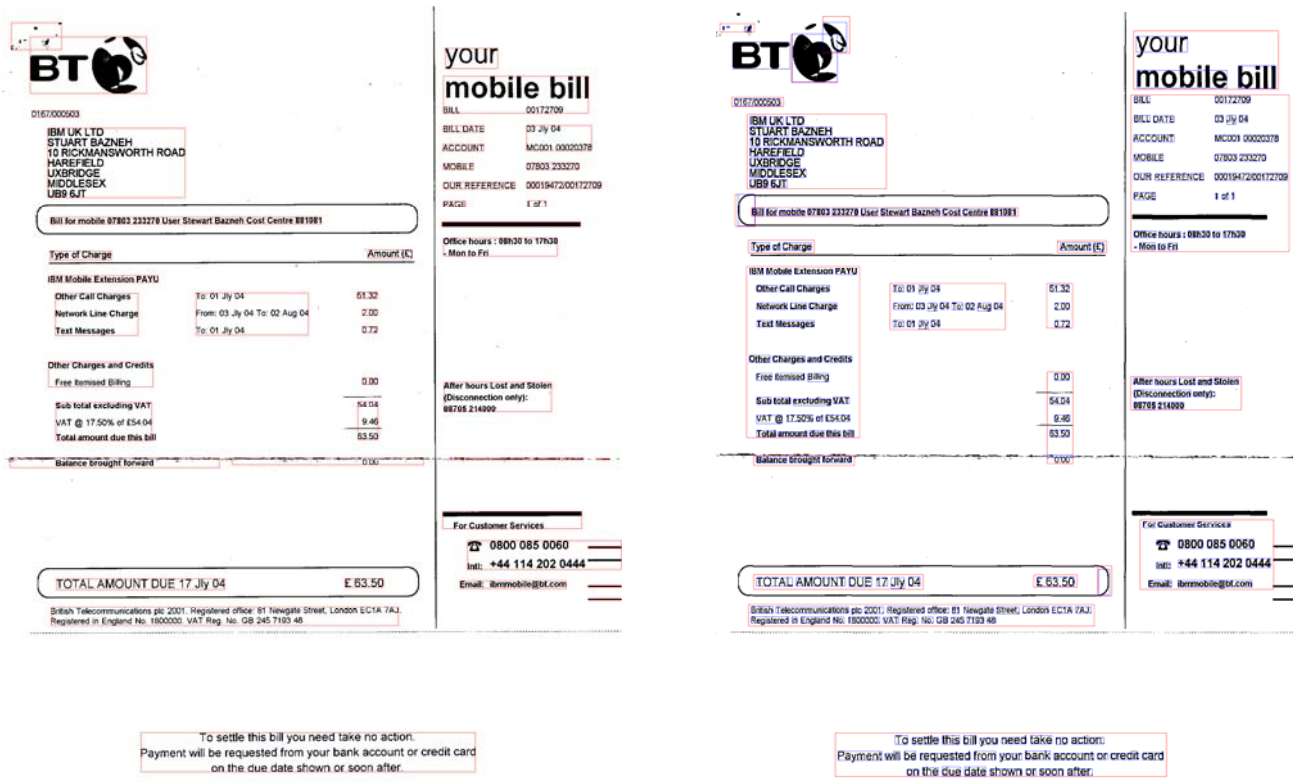

Figure 4: Page segmentation results by the Docstrum algorithm (left) and by the OCR engine (right).

OCR for page segmentation avoids the tedious step of training the free parameters involved in a stand-alone algorithm. Packaged OCR products provides a convenient black-box solution, in which the engine parameters have been tuned for optimal performance over large collections of documents. At the post-processing stage, the OCR engine can use preliminary recognition results to further improve page segmentation in an iterative fashion. Additional information, including consistency in the font style and spatial alignment of segmented regions, help improve overall page segmentation performance, and tend to produce results that are more structurally meaningful, even if the input image is heavily degraded. This is also observed in past empirical studies [28], where representative page segmentation algorithms are evaluated against built-in page segmentation functions provided by a few early OCR products.

3.3 Feature Selection Issues

Optical character recognition on machine-printed characters has emerged as an industrial-strength technology since its phenomenal advances from early 1990s. However, OCR accuracy today is still far from that of a second-grade child in many aspects [29]. It is important to put the strength and weakness of OCR technology in perspective, and understand the factors involved that have significant impact on its performance. These insights provide useful guidelines for selecting feature sets that can be extracted relatively reliably given the practical constraints imposed by a targeted application.

The most successful application domain of OCR technology to date is on machine-printed characters. Over the last decade, the acceptance rates of form readers on hand-printed digits and constrained alphanumeric fields have also risen significantly. The relatively low recognition errors in these constrained domains are a reflection of the complexities involved in classifying a novel pattern under such limited variation in the data set [30]. In contrast, recognition of unconstrained off-line human handwriting and multi-lingual recognition among a variety of scripts are much more challenging problems, and they still remain as active research frontiers.

The accumulated imaging degradations have a significant impact on OCR performance. Typical imaging defects in the printing process include blotchy characters caused by dot-matrix printer ribbons, and faint impressions as a result of worn ribbons and printer cartridges. The scanning process introduces various imperfections of its own. Digital scanning involves sampling both horizontally and vertically on the image grid. Desirable sampling rates by OCR are beyond 300 dots per inch (dpi). Although commercial packages can work at as low as 150 dpi by interpolating a low-resolution image to the preferred dpi, this generally leads to significant increase in recognition errors. Significant image degradation also occurs when storing an image in binary format by applying thresholds to separate foreground content from the background. Using gray-scale and color scans of the image captures more detailed information for pattern recognition and reduces the error rate. Most high-end MFDs today provide these functions.

3.4 Feature Set

We use a rich combination of page layout and linguistic features. The feature extraction process can be viewed as a set of binary-valued functions defined on the appropriate feature space that output either 1 or 0 based on the presence or absence of the corresponding feature. The conditional nature of CRFs enables effective learning from these discrete-valued features with different granularity and multiple modalities, which may have extremely complex joint probabilities.

3.4.1 Layout Features

As shown in Fig. 4, noise speckles and graphic elements in the document, including logos, lines, and region borders, may not be reliably classified and segmented, and thus be wrongly fed to the machine-printed text recognizer. We simply discard those segmented text regions, in which majority of text is un-recognizable or suspicious. We use the following collection of layout-related features extracted by the OCR engine on each segmented region:

- Variation of font size *and* font face within the region
- Presence of the largest font on the entire page
- Presence of bold font face in the region
- Whether the text block is horizontally aligned to the center
- Whether the text block horizontally aligned to the left

3.4.2 Text Features

Word tokens that are logically or semantically related to a NE are very useful and relatively robust features for extracting the NE. In fact, a common technique for character-level error correction used by current OCR systems is to explicitly make use of context at word level, by choosing a common letter n-gram over a rare one [29]. This gives improvement in recognition performance at word level even if the quality of the image is very poor. We organize word tokens into equivalent groups. For example, "Inc." and "Companies" are grouped together. The following text features are also used:

- Capitalization of words
- Mixed cases
- Frequent appearance of digit characters (0-9)
- Presence of special characters (/, –, #, -, *, $, £)
- Presence of special patterns ('s)

3.4.3 Named Entity Features

The orthogonality between NEs can be effective features for inference. A region of text containing a credit card number is less likely to contain the name of the merchant. We use the set of orthogonally related NEs as features. These include addresses, phone numbers, credit card numbers, dates, and monetary amounts.

3.5 Region Labels

We use a compact set of labels to categorize the ordered list of regions obtained by page segmentation. This is based on our observation that context change along the sequence of regions is frequent, making the inference of label more effective among neighboring regions. For extracting the merchant, we use three labels of regions, NON_DATA, MERCHANT_DATA, and TRANS_DATA. NON_DATA represents region that does not contain details of a transaction or any association with a merchant. MERCHANT_DATA denotes a region that contains a merchant. TRANS_DATA region includes details of a transaction.

3.6 Relevant NE Extraction with CRFs

A *conditional random field* can be viewed as an undirected graphical model, and be used to compute the conditional probability of labels on designated output nodes $\boldsymbol{Y}$, when globally conditioned on $\boldsymbol{X}$, the random variable representing observation sequences. In a discriminative CRF framework, we construct a conditional model $p(\boldsymbol{y} \mid \boldsymbol{x})$ from paired observation and labeled sequences, and do not explicitly model the marginal $p(\boldsymbol{x})$.

We work with CRFs with a linear chain structure. Given an instance of observation sequence $\boldsymbol{x}$, the probability of a particular label sequence $\boldsymbol{y}$ is defined in [8] as

$$p(\boldsymbol{y} \mid \boldsymbol{x}) \propto \exp\Big(\sum_j \lambda_j f_j(y_{i-1}, y_i, \boldsymbol{x}, i) + \sum_k \mu_k f_k(y_i, \boldsymbol{x}, i)\Big), \quad (1)$$

where $f_k(y_i, \boldsymbol{x}, i)$ is a state feature function of the label at position i and the observation sequence; $f_j(y_{i-1}, y_i, \boldsymbol{x}, i)$ is a transition feature function of the entire observation sequence and labels at positions i and $i-1$. More compactly, the probability of a label sequence $\boldsymbol{y}$ given the observation sequence $\boldsymbol{x}$ is given by

$$p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\lambda}) = \frac{1}{Z(\boldsymbol{x})} \exp\Big(\sum_j \lambda_j F_j(\boldsymbol{y}, \boldsymbol{x})\Big), \quad (2)$$

where $Z(\boldsymbol{x})$ is a normalization factor and

$$F_j(\boldsymbol{y}, \boldsymbol{x}) = \sum_i f_j(y_{i-1}, y_i, \boldsymbol{x}, i).$$

Assuming the training data $\{\boldsymbol{x}^k, \boldsymbol{y}^k\}$ are independently and identically distributed, the product of Equation (2) over all training sequences as a function of the parameters $\boldsymbol{\lambda}$ is the likelihood function. Maximum likelihood training chooses parameters values such that the log-likelihood is maximized. For a CRF, the log-likelihood is a concave function, guaranteeing convergence to the global maximum.

$$L_\theta(\boldsymbol{\lambda}) = \sum_k \left[\log \frac{1}{Z(\boldsymbol{x}^k)} + \sum_j F_j(\boldsymbol{y}^k, \boldsymbol{x}^k) \right]. \quad (3)$$

Likelihood maximization can be performed efficiently using a quasi-Newton method, such as L-BFGS [31]. This method approximates the second derivative of the likelihood by keeping a running, finite window of previous first-derivatives. L-BFGS can simply be treated as a black-box optimization procedure, requiring only the first derivative of the function to be optimized.

Let $\boldsymbol{y}^j$ be the state path up to position T on instance j of the labeled training sequence. The first-derivative of the log-likelihood function is given by

$$\frac{dL_\theta}{d\theta} = \sum_{j=1}^{N}\left(\sum_{i=1}^{T}\frac{dF_\theta(y_t^j, y_{i-1}^j, \boldsymbol{x}^j)}{d\theta} - \sum_{\boldsymbol{x}} p(\boldsymbol{y} \mid \boldsymbol{x}^j)\sum_{i=1}^{T}\frac{dF_\theta(y_i, y_{i-1}, \boldsymbol{x}^j)}{d\theta}\right) \quad (4)$$

Intuitively, when the state paths chosen by the CRF parameters match the state paths from the labeled sequence, the derivative given in Equation (4) becomes zero.

4. RESULTS AND DISCUSSION

4.1 Datasets

We used two large real-world receipt collections provided by IBM World Wide Reimbursement Center for training and testing, which contain binary scanned receipt images from both IBM internal business units and IBM Global Services customers. These two collections are realistic as these paper receipts were gathered and scanned over time using a variety of equipment. Characteristics of the two datasets are summarized in Table 1.

Table 1: The two real-world receipt image collections.

	Collection 1	Collection 2
Total images	145	283
Number of characters	71316	522320
Character error rate	6.03%	9.48%
Image resolution	200-300 dpi	150-200 dpi
Country of origin	US	UK

We used the first two fifth from each dataset for training. Groundtruth labels were created by first running a rule-based heuristic on the sequence of segmented regions. Human judgment was then employed to correct mistakes in the heuristic labeling.

4.2 Evaluation and Discussion

We use the precision-recall metrics to evaluate performance. From a practical point of view, we want our metrics to provide unbiased estimate of performance realistically achievable by including the effect of recognition errors. Throughout our evaluation, we define *recall* as the ratio of the number of NEs correctly extracted to the number of NEs that are physically present in the collection. *Precision* is the ratio of the number of NEs correctly extracted divided by the total number of NEs extracted in the category. The *F-Measure* (or F_1 Measure) is computed by ($2 \times Precision \times Recall$) / ($Precision + Recall$) [32].

Tables 2 and 3 summarize the overall NE extraction performances on the two collections of receipt images, respectively. On both datasets, extraction of merchant using CRF outperformed its heuristic counterpart. This is an encouraging result, since improvements on a rule-based system requires constant changes to the code base, while improvements on the CRF system generally require only defining new features that can be automatically extracted. In fact, we observed improvement in performance after modifying the word token sets to reflect locale difference of the two datasets.

Table 2: Performance of NE extraction on collection 1.

	Precision	Recall	F-Measure
Date	100.0	85.2	92.0
Credit Card #	98.8	76.5	86.2
Expense total	100.0	97.7	98.8
Phone #	95.3	78.1	85.8
Address	95.4	82.7	88.6
Merchant (by heuristic)	63.2	56.8	59.8
Merchant (by CRF)	73.8	70.5	72.1

The impact of recognition errors on rule-based NE extraction approaches is evident. Almost all the errors made by the heuristic on simple NEs were caused by errors in the text. Image scan on a higher-resolution device or a device that supports gray-scale or color formats can levitate this problem by effectively containing the recognition errors. In addition, a paper document goes through image capturing (scanning) only once in our system in contrast to some receipts in our evaluation datasets, which are second-generation copies (e.g. scanned image of a previously faxed document). The reported figures provide a realistic, and somehow conservative estimate of the performance in a field deployment, as digital printing technologies have significantly improved over the last decade.

Table 3: Performance of NE extraction on collection 2.

	Precision	Recall	F-Measure
Date	96.3	78.2	86.3
Credit Card #	92.4	69.5	79.3
Expense total	89.1	91.7	90.4
Phone #	84.8	66.1	74.3
Address	82.3	58.6	68.5
Merchant (by heuristic)	58.5	49.7	53.7
Merchant (by CRF)	67.2	62.9	65.0

NE extraction approach using CRF is shown to be more robust to recognition errors, as it jointly looks at collection of features. For instance, the 9.48% character error rate on collection 2 translates to a 39.2% word level error rate for a five-letter word. Nevertheless, we achieved 65.0% F-score on merchant using CRF on this dataset.

At the current level of NE extraction performance, some human assists are still required in a live expense reimbursement system. Of course, manual efforts are necessary for handling practical issues that are unrelated to data mining performance, such as instances of lost or illegible receipts and incorrect submissions. We will present a transformed expense reporting workflow that greatly relaxes these requirements.

4.3 Practical Impacts

Paper document is the most common form of information-conveying vehicle. Humans have developed exceptional capability to extract information from paper documents, even though the text content is un-structured. However, when they are scanned into image form and processed by machines, it becomes very difficult in general to index and mine, given the high-volume, complex and the heterogeneous nature of data in many practical applications. Extraction of relevant metadata from semi- and un-structured paper documents is fundamental to a range of novel and important applications in this area. Such challenges have been posed at SIGKDD 2006 conference by a panel of data mining experts [33].

We have presented an effective multimedia mining approach that aims to cross the semantic gap between un-structured image data and semantics by jointly learning page layout and linguistic features. Such multimedia mining capability can be an essential technical component in many application domains that require image-based document indexing and retrieval.

From an application point of view, extracting transaction-related named entities from expense receipts provides a good starting point for a range of data mining applications for an organization. For example, we can start to ask a series of questions: How much are employees of an organization spending on a particular expense item? What expense items are redundant? How can an organization analyze these spending patterns and optimize them under practical constraints?

Over the years, we have seen an increasing number of domain-specific applications that address various needs of document image processing at a particular step of a business process. However, integrating these loosely coupled applications proves to be challenging, since their underlying system architectures are very much ad hoc in nature. With the emergence of MFDs as a consolidated electronic imaging platform in the office, a common set of document import, transport, and storage facilities becomes very important. Our system provides such a flexible and scalable document management infrastructure on which multi-type paper documents can be aggregated in digit form and processed intelligently based on their process requirements.

4.4 Experiences and Lessons Learned

Throughout this project, we have been working closely with teams from IBM Global Services and IBM Printing Systems. Collaboration across business units offers a fast route of technology transfer from corporate research labs to product and service offerings that make real business impact. Working with subject experts from these divisions also provides us with valuable insight into business processes. In fact, many features related to automation of auditing procedures came from discussions with these subject experts.

As Thearling [34] has pointed out, the core data mining algorithms are currently a small part of the overall application, being perhaps 10% of a larger part, which itself is only 10% of the whole. Finding the killer application for applying data mining algorithms proves to be a key success factor to this project. Our solution also leverages on the huge market potential created by the technology shift to consolidated digital MFD platforms in the office environment.

4.5 Future Work

Evaluation on real-world datasets has shown a very promising perspective on the field deployment of our fully automated expense reimbursement solution. More importantly, this system can evolve into a multi-document-type platform, which integrates variety of business processes that traditionally have been heavily involved with paper documents.

One direction we are working towards is to streamline the expense reporting steps and make the entire process more human friendly. Our vision is to merge the web-based expense reporting seamlessly with electronic submission. The new workflow outlined below has the advantage that it involves well-informed user decision when making corrections. In addition, with the presence of receipt image, immediate expense claim approval from management is made possible for expenses that exceed certain amount or in special categories, such as procurement and tuition reimbursement.

- Submit receipt documents electronically
- Open the email notification sent by the system once the submitted receipt has been processed, and click on the link in the email to a web-based application
- With the submitted receipt image side by side, make corrections to the fields on the web form, which have been mostly pre-filled up using the automatically extracted NEs
- Submit the completed reimbursement claim

With the current trend of globalization, busines travels across countries are frequent. Another major feature expected to be integrated into our solution is the *multi-language support* for global deployment. It includes an additional language identification component that determines the language set of a document before running the recognizer of the language.

5. CONCLUSION

We have described an emerging data mining application for corporate expense reimbursement. At the system level, our end-to-end solution involves technical innovations in three areas: (1) a generalized paper-free framework for capturing, transporting, and storing paper documents in digital image form; (2) an unconstrained data mining approach to extracting relevant named entities from un-structured document images; (3) automation of auditing procedures for minimizing manual efforts. We are actively collaborating with IBM Global Services and IBM Printing Systems for pilot testing our solution.

Our contributions to the practice of data mining are (1) a unified model-based approach to jointly solving the named entity extraction and question answering aspects of the problem using conditional random fields (CRFs); (2) a formal framework for efficient probabilistic inference by collectively learning the contextual dependencies from both page layout and linguistic features; and (3) evaluation of our data mining approach using un-structured document images. Our approach is general and can be applied to image-based information retrieval problems in broader application domains, where the text itself reveals limited structural information.

6. REFERENCES

[1] Mori, S., Suen, C. Y., and Yamamoto, K. (1992). Historical review of OCR research and development. In *Proc. of the IEEE*, 80(7), 1029–1058.

[2] Callan, J., Kantor, P., Grossman, D. A. (2002). Information retrieval and OCR: from converting content to grasping meaning, *SIGIR Forum*, 36(2), Tampere, Finland.

[3] Chinchor, N. (1998). Named entity task definition. In *Proc. of the Message Understanding Conference*.

[4] Borthwick, A., Sterling, J., Agichtein, E., and Grishman, R. (1998). Exploiting diverse knowledge sources via maximum entropy in named entity recognition. *WVLC 1998*.

[5] Lawrence, S., Giles, C. L., and Bollacker, K. (1999). Digital libraries and autonomous citation indexing. *IEEE Computer*, 32(6), 67–71.

[6] Krishna, V. and Srinivasan, S. (2006). *System, method, and service for automatically and dynamically composing document management applications*. US patent granted, US2004000980716.

[7] Krishna, V. and Srinivasan, S. (2006). *System and method for defining and generating document management applications for model-driven document management*. US patent granted, US20060253490A1.

[8] Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. *ICML 2001*.

[9] McCallum, A., Nigam, K., Rennie, J., and Seymore, K. (2000). Automating the construction of internet portals with machine learning. *Information Retrieval*. 3:127–163.

[10] Bunescu, R., Ge, R., Mooney, R. J., Marcotte, E., and Ramani, A. K. (2002). Extracting gene and protein names from biomedical abstracts. Unpublished Technical Note, http://www.cs.utexas.edu/users/ml/publication/ie.html.

[11] Humphreys, K., Demetriou, G., and Gaizauskas, R. (2000). Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures. *PSB 2000*, 502–513.

[12] Etzioni, O., Cafarella, M., Downey, D., Popescu, A.-M., Shaked, T., Soderland, S., Weld, D. S. and Yates, A. (2005). Unsupervised named-entity extraction from the web: an experimental study. *Artificial Intelligence*, 165, 91–134.

[13] Fukuda, K.,Tsunoda, T.,Tamura, A., and Takagi, T. (1998). Toward information extraction: identifying protein names from biological papers. *PSB 1998*, 707–718.

[14] Narayanaswamy, M., Ravikumar, K. E., and Vijay-Shanker, K. (2003). A biological named entity recognizer. *PSB 2003*, 427–438.

[15] Carrerars, X., Marques, L. and Padro, L. (2002). Named entity extraction using Adaboost. *CoNLL 2002*, 167–170.

[16] Florian, R., Ittycheriah, A., Jing, H., and Zhang, T. (2003). Named entity recognition through classifier combination. *CoNLL 2003*, 168–171.

[17] Collins, M. (2002). Discriminative training methods for Hidden Markov Models: Theory and experiments with perceptron algorithms. *EMNLP 2002*, 10, 1–8.

[18] McCallum, A., Freitag, D., and Pereira, F. (2000). Maximum entropy Markov models for information extraction and segmentation. *ICML 2000*, 591–598.

[19] Wallach, H. M. (2004). *Conditional Random Fields: An Introduction*. Technical Report, University of Pennsylvania.

[20] Pinto, D., McCallum, A., Wei, X., and Croft, W.B. (2003). Table extraction using conditional random fields. *SIGIR 2003*, 235–242.

[21] Sha, F. and Pereira, F. (2003). Shallow parsing with conditional random fields. *HLT-NAACL 2003*, 1, 134–141.

[22] Sarawagi, S. and Cohen, W. (2004). Semi-markov conditional random fields for information extraction. NIPS.

[23] He, X., Zemel, R., and Carreira-Perpinan, M. (2004). Multiscale conditional random fields for image labeling. *CVPR 2004*, 2, 695–702.

[24] Smimchisescu, C., Kanaujia, A., Li Z. and Metaxus D. (2005). Conditional models for contextual human motion recognition. *ICCV 2005*, 2, 1808–1815.

[25] Cohen, W. and Sarawagi, S. (2004). Exploiting dictionaries in named entity extraction: Combining semi-markov extraction processes and data integration methods. *KDD 2004*, 89–98.

[26] O'Gorman, L. (1993). The document spectrum for page layout analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15, 1162–1173.

[27] Kise, K., Sato, A., and Iwata, M. (1998). Segmentation of page images using the area Voronoi diagram. *Computer Vision and Image Understanding*, 70, 370–382.

[28] Mao, S. and Kanungo, T. (2001). Empirical performance evaluation methodology and its application to page segmentation algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(3), 242–256.

[29] Rice, S.V., Nagy, G., and Nartker, T.A., (1999). *Optical Character Recognition: An Illustrated Guide to the Frontier*. Kluwer Academic Publishers.

[30] Duda, R.O., Hart, P.E., and Stork, D.G. (2000). *Pattern Classification* (*2nd Edition*). Wiley-Interscience.

[31] Byrd, R.H., Nocedal, J., and Schnabel, R.B. (1994). Representations of quasi-Newton matrices and their use in limited memory methods. *Mathematical Programming*, 63:129–156.

[32] Ng, H.T., Kim, C.Y., and Koo, J.L.T. (1999). Learning to recognize tables in free text. *ACL 1999*, 443–450.

[33] Piatetsky-Shapiro, G., Djeraba, C., Getoor, L., Grossman, R., Feldman, R., and Zaki, M. (2006). Is there a grand challenge or X-prize for data mining? *KDD 2006*, 954–956.

[34] Thearling, K. (1998). Some thoughts on the current state of data mining software applications. *KDD 1998 Workshop: Keys to the Commercial Success of Data Mining*.