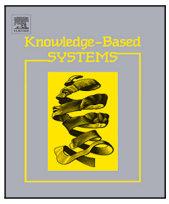
Knowledge-Based Systems 311 (2025) 113026

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Corporate relation extraction for the construction of knowledge-bases against tax fraud

Inigo Lopez-Gazpio [a],*, Laura Baselga-Pascual [b], Aitor Garmendia-Lazcano [b]

[a] HiTZ Basque Center for Language Technology - Ixa NLP Group. University of the Basque Country UPV/EHU, Donostia, Spain
[b] Department of Finance and Economics, Deusto Business School. University of Deusto, Donostia, Spain

ARTICLE INFO

Dataset link: https://www.boe.es/diario_borme/

Keywords:
Computer Vision
Information extraction
Knowledge-base generation
Natural Language Processing
Tax fraud investigation

ABSTRACT

Tax fraud is a criminal activity that entails significant losses for governments. Due to its clandestine nature, it is difficult to reliably estimate the amount of taxes evaded. To fight tax fraud, this investigation details the construction and evaluation of a corporate relation extraction system designed to access an unstructured knowledge-base and extract corporate relations for further validation. The system was developed in response to a need raised by the Treasury and Finance Department of the Provincial Council of Gipuzkoa (Spain). It follows a waterfall architecture that integrates Natural Language Processing (NLP) and Computer Vision (CV) components, including web scraping, optical character recognition, syntactic parsing, and information extraction. The proposed system produces a relational knowledge-base with structured data representing 23 types of corporate operations published in the Official Gazette of the Commercial Registry (e.g., incorporation of companies, terminations, capital increases and reductions, mergers and takeovers, etc.), allowing for comparison with the fiscal information available in the tax agency. Facilitating such comparison across distinct sources is key to identifying discrepancies that might be indicators of tax fraud.

1. Introduction

Tax fraud is defined as the illegal practice of tax evasion carried out by individuals and companies. Generally, tax evasion involves the deliberate manipulation of information submitted to the tax authorities. Some examples of tax evasion include under-declaring income or profits, concealing taxable assets, or overstating tax deductions [1,2].

Given the clandestine nature of tax fraud, it is difficult to reliably estimate the loss to governments' tax revenues. Some authors, such as [3], estimated that the money hidden from tax scrutiny globally ranged from $21 to $32 trillion in 2010, equivalent to the combined Gross Domestic Product (GDP) of the United States, Japan, and Germany. Other tax evasion studies, such as [4], estimated that the income loss due to tax evasion may reach 5% of global GDP.

The recent trend among regional, national, and international tax authorities to combat tax fraud is reflected in the increased use of the collocation "tax fraud" in publications indexed in the Journal Citation Report (JCR). Fig. 1 shows the evolution of tax fraud-related research. As can be noted from the figure, the number of published manuscripts in the area has drastically increased over the last twenty years. The growing interest from academia in studying tax fraud reveals the willingness of governments and tax authorities to address this problem that depletes public coffers. Moreover, the phenomenon of tax fraud has been amplified by globalization and the increasing digitalization of the economy.

A first approach to combat tax fraud consists of evaluating the correspondence between different data sources. For instance, matching data from tax agencies with other sources allows one to test for misinformation or inconsistencies. However, to the best of our knowledge, the detailed construction and evaluation of a relation extraction tool that enables subsequent tax validation has never been documented in the State-Of-the-Art (SOtA). The availability of such a resource can potentially streamline the analysis of large amounts of information on business operations, such as the incorporation of companies, dismissals, capital increases and reductions, dissolutions, takeovers, mergers, and divisions.

Determining the extent to which automated tools can help combat tax evasion is a major concern in a digitalized era in which multiple corporate operations are recorded in the cloud. To this end, this work responds to the need expressed by the Treasury and Finance Department of the Provincial Council of Gipuzkoa (Spain) to create and evaluate a relation extraction tool to process unstructured data from the Official Gazette of the Commercial Registry (BORME from its Spanish translation *'Boletín Oficial del Registro Mercantil'*) to evaluate the feasibility of employing an automated extraction tool to cross-reference

* Corresponding author.
E-mail addresses: inigo.lopez@ehu.eus (I. Lopez-Gazpio), lbaselga@deusto.es (L. Baselga-Pascual), aitor.garmendia@deusto.es (A. Garmendia-Lazcano).

https://doi.org/10.1016/j.knosys.2025.113026
Received 9 December 2021; Received in revised form 30 October 2024; Accepted 11 January 2025
Available online 19 January 2025

Fig. 1. Number of indexed articles published in the Journal Citation Reports with the keyword "Tax fraud" between 1996 and 2020.

it with the information provided by the tax agency to detect possible inconsistencies that might lead to tax fraud. Our relation extraction tool allows for the consistent and accurate extraction of large volumes of information published online for easier subsequent investigation, helping to identify possible discrepancies that could indicate potential tax fraud.

This investigation details the development and evaluation of a relation extraction system based on NLP and CV technology that is capable of automating the extraction of relations from a set of corporate operations published in the BORME, and building a structured knowledge-base. This knowledge-base can bring together a large amount of information on different corporate operations included in the BORME (i.e., incorporation of companies, capital increases and reductions, dissolutions, mergers, absorptions and split-offs, among others) and enable the Gipuzkoan tax agency to quickly match their own data with the public information contained in the Official Gazette of the Commercial Registry.

Thus, the primary contribution of this work to the fight against tax fraud is the development of a tool that systematically transforms the unstructured and expansive data found in the BORME into a structured knowledge-base, specifically modeling corporate activities. Tax authorities, such as the Provincial Council of Gipuzkoa, currently face challenges in cross-referencing corporate data with fiscal records due to the fragmented, unstructured, and voluminous nature of available data sources. This makes the identification of discrepancies and potential tax fraud labor-intensive and inefficient. The tool we propose addresses these challenges by automating the extraction and organization of relevant corporate relations, allowing for more efficient and scalable analysis. Furthermore, this paper thoroughly evaluates the relation extraction tool, verifying both the correspondence and accuracy of the extracted relations and attributes, and demonstrating the system's feasibility and performance in a real-world setting.

The paper is structured as follows: Section 2 analyzes previous investigations and technological resources from the NLP and CV fields related to the extraction of information that can be used for the inspection of tax fraud. Section 3 describes the main source of information being utilized and its unstructured nature. Section 4 outlines the proposal and construction of the relation extraction system, detailing its components and the set of relations extracted. Section 5 presents the experiments performed to evaluate the extraction system, and, finally, Section 6 concludes the research and outlines future work.

2. Related work

Tax fraud mitigation represents a complex problem and ongoing challenge. Systems constructed for the investigation of tax fraud aim to monitor user behavior, entities, or companies to gather data in order to model profiles and behaviors that might constitute fraudulent actions. Earliest studies addressing the detection of 'a posteriori' fraud in the fiscal and insurance domains are described by [5]. The authors report that planning adequate audit strategies is key to detecting tax evasion and fraudulent claims, and that the required audit strategy can be supported by a feature-based classification [6]. The investigation is supported by a case study in which the authors show the trade-off between maximizing audit benefits and minimizing audit costs. More recent investigations on the problems of the shadow economy, tax evasion, and tax avoidance, such as the work of [7], focus on the construction of an intelligent system [8] that is able to monitor the behavior of accounting information systems. This work also focuses on the rapid growth of Information and Communications Technologies (ICT) and how effective behavioral monitoring systems should account for this revolution and the large volume of data contained within online transactions. The works of [9–12] are also relevant in the field of big data analytics, in which the usage of large knowledge-bases containing corporate information gathered from diverse sources and updated at different intervals is key for the investigation, identification, and comparison of tax fraudulent actions.

In order to gather valuable data and investigate tax fraud, several data mining techniques have been used in the SOtA. The works of [13,14], and [15] highlight the following NLP and CV technologies: (i) optical character recognition [16], (ii) machine translation [17], (iii) automatic summarization [18], (iv) information extraction [19], (v) named entity recognition [20], and (vi) relation extraction [21].

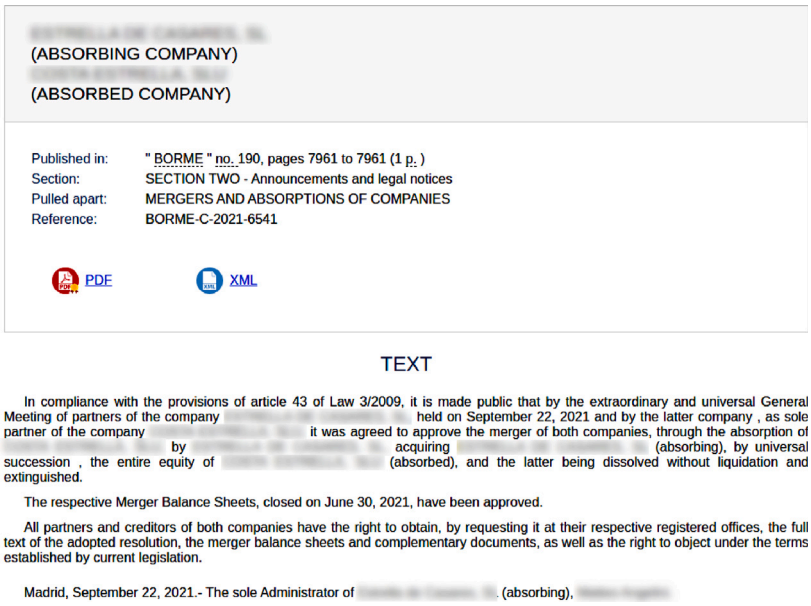
The work of [22] describes how artificial intelligence experts worked alongside specialists from the Italian Revenue Agency to build a rule-based audit system created from data collected using some of the mentioned information extraction techniques. In an initial scenario that estimated tax fraud in Italy at around 3%–10% of the country's GDP, the authors sought to recover large amounts of resources by optimizing company audits (well-targeted audits). The investigation concluded that planning adequate audit strategies through expert systems based on knowledge-bases is a practical and effective approach compared to classical approaches. [23] also focuses on data mining and information extraction techniques. In their work, the authors describe how any tax, such as the Value Added Tax (VAT), is susceptible to fraud and evasion, and for this task, several information extraction techniques (such as rules based on finite automaton cascades [24] and statistical NLP techniques [25]) are employed to model the behavior of tax evaders. In this endeavor, converting unstructured text into a computationally manageable or human-friendly format has been one of the keys to effective subsequent analysis. From the very first investigations of tax fraud, [5] highlighted the importance of building a knowledge-base containing corporate operations as a first step towards the investigation of tax evasion, as it enables further behavioral modeling, analysis, and comparison.

3. Information source and the unstructured nature of the BORME

Access to a large knowledge-base of corporate information is a fundamental component for the posterior investigation of tax fraud. We now describe the BORME,[1] which is the main source utilized for our extraction tool.

Due to the unstructured nature of the BORME (See Fig. 2), the Treasury and Finance Department of the Provincial Council of Gipuzkoa requires a relation extraction tool to convert the information it contains into a structured knowledge-base. BORME is the main instrument for the publication of information from the Commercial Registry in Spain. Legal and economic data of the companies registered in this bulletin are published on a daily basis, comprising a large volume of information that is continuously growing. All the contents of the digital edition of

[1] https://www.boe.es/diario_borme/.



[redacted]
(ABSORBING COMPANY)
[redacted]
(ABSORBED COMPANY)

Published in: " BORME " no. 190, pages 7961 to 7961 (1 p.)
Section: SECTION TWO - Announcements and legal notices
Pulled apart: MERGERS AND ABSORPTIONS OF COMPANIES
Reference: BORME-C-2021-6541

PDF XML

TEXT

In compliance with the provisions of article 43 of Law 3/2009, it is made public that by the extraordinary and universal General Meeting of partners of the company [redacted] held on September 22, 2021 and by the latter company , as sole partner of the company [redacted] it was agreed to approve the merger of both companies, through the absorption of [redacted] by [redacted] acquiring [redacted] (absorbing), by universal succession , the entire equity of [redacted] (absorbed), and the latter being dissolved without liquidation and extinguished.

The respective Merger Balance Sheets, closed on June 30, 2021, have been approved.

All partners and creditors of both companies have the right to obtain, by requesting it at their respective registered offices, the full text of the adopted resolution, the merger balance sheets and complementary documents, as well as the right to object under the terms established by current legislation.

Madrid, September 22, 2021.- The sole Administrator of [redacted] (absorbing), [redacted]

Fig. 2. Example extracted from the second section of the BORME showcasing an anonymized absorption relation between two corporations. The example has been translated from Spanish.

the BORME are publicly accessible on the internet, offered in different formats and organized into two sections:

(i) The First Section of the BORME, called *'Entrepreneurs'*, contains information on entrepreneurs, registered acts, legal acts, company incorporations, appointments, re-elections, and dismissals of administrators, to mention some of the most notable operations.

(ii) The Second Section of the BORME, called *'Announcements and legal notices'*, contains notifications, communications, and legal notices regarding registered corporations. This section is subdivided into a series of subsections, among which we highlight: announcements from the Central Commercial Registry, balance sheets, calls for meetings, declarations of insolvency, dissolutions of companies, mergers and takeovers, assignments, capital increases, capital reductions, and other announcements and legal notices.

The legal acts contained in the First Section are organized by province, with each block of data following the postal codes of the respective provinces. In contrast, the publications in the Second Section follow an alphabetical order by company name. We define all the corporate relations targeted for exploitation in Section 4.3.

4. Construction of the system

This section describes the implementation of the system, detailing the techniques and resources that have been used to build the extraction tool. The main objective of the relation extraction tool is to automate the task of identifying and gathering relevant relations from the BORME, processing them into structured relations and building a knowledge-base with structured data. For the task, the system makes use of several NLP and CV techniques, such as web scraping frameworks, OCR (Optical Character Recognition), tokenizers, text normalization, named entity extractors, morphosyntactic tagging, and finite automata. One of the major challenges concerning the relation extraction tool consists of the unstructured nature of the data contained in the BORME, from which we aim to collect certain corporate relations.

Fig. 3 illustrates the main four components of the extraction system and their interactions using a Unified Modelling Language (UML) component diagram. As shown in the diagram, the main components of the system are as follows: (i) the web scraping component, (ii) the OCR component, (iii) the relation extraction component, and (iv) the persistence component. We briefly describe these components below.

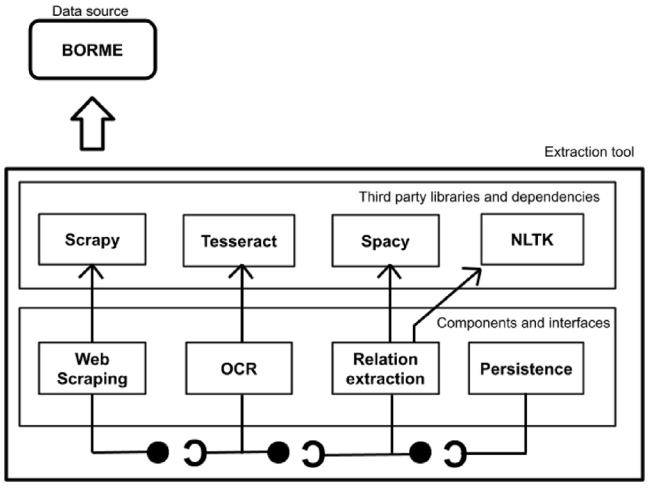


Fig. 3. The UML diagram illustrates the main components of the proposed system. From left to right, the components displayed are: (i) the web scraping component, (ii) the OCR component, (iii) the corporate relation extraction component, and, finally, (iv) the persistence component.

(i) The web scraping component is responsible for the web crawler that handles a list of target web addresses from which to extract unstructured data. Web scraping frameworks operate by simulating a web browser to read HyperText Markup Language (HTML) code and extract the relevant content. The content is returned as unstructured data for further processing by the subsequent components. An example of the content of the HTML code is shown in Fig. 2. Additionally, the web scraping component is able to download Portable Document Format (PDF) files associated with the target web address.

(ii) The OCR component is used to extract unstructured data from the PDF files. This process aims to digitize text using artificial intelligence techniques [26]. In the particular case of the BORME website, the PDF files containing corporate relations are composed of a sequence of images that need to be processed in order to obtain the underlying text. We use the OCR component to extract the text contained in the images through CV techniques.

(iii) The corporate relation extraction component is the main component of the information extraction tool, as it encompasses the set of NLP techniques used to identify and extract corporate relations. This component consists of a series of subcomponents that run in a cascade and process the unstructured text originating from HTML code or PDF files until it is transformed into a structured set of defined corporate relations. To process and give a defined structure to the information, we tokenize and lemmatize the text to achieve a standardized format by applying morphosyntactic taggers. Then, we extract named entities to identify people, organizations, places, dates, and amounts. Finally, corporate relations are identified and extracted using finite automata.

(iv) The persistence component stores the extracted relations, which comprise the structured knowledge-base, as agreed with the tax agency, for later analysis by the agency's technicians.

All the components described above are organized in such a way that each of them implements a service with single responsibility (see Fig. 3). The objective of such a design is to comply with good software development practices, more commonly known as the *SOLID design principles* [27]. We now further describe the details of each component.

4.1. The web scraping component

The extraction of raw HTML and PDF documents is handled largely by a web scraping framework named *Scrapy*. Scrapy[2] is an open-source web scraping and web crawling framework written in Python that allows the implementation of web spiders capable of automating web content extraction. To launch Scrapy, we first automate the creation of target Uniform Resource Locators (URLs) following the format of BORME to indicate the target time window subject to exploitation, and provide the HTML sections to extract through XPath queries. URLs are defined as strings that follow the format, https://www.boe.es/borme/dias/yyyy/mm/dd where *y* stands for the year specified using 4 digits, *m* stands for the 2-digit month, and *d* stands for the 2-digit day. Once the target address is set, the content for extraction is specified using XPath queries that indicate relevant segments of the Document Object Model (DOM). XPath is a well-known technique for information extraction in the area of NLP [28].

4.2. The OCR component

The conversion of images contained in PDF files into unstructured text is performed through the OCR tool Tesseract [16], and its interface over Python, known as pytesseract.[3] Pytesseract is a low-level interface that allows the invocation of the neural networks contained in the OCR tool in a parameterized way. In its most recent version,[4] Tesseract applies a cascade of algorithms to perform character recognition, including the detection of blocks to identify the spatial structure of the document [29] and the identification of isolated lines and characters by means of convolutional neural networks (CNNs) or adaptive algorithms [16]. Investigations such as those carried out by [30,31] have confirmed Tesseract as a SOtA OCR tool compared to other OCR systems, highlighting the good performance of the tool under different types of resolution, brightness, or font quality conditions.

The detection of blocks and identification of spatial structure are performed by an iterative block detection algorithm based on the work of [29], in which the authors describe the inclusion of a measure to gradually evaluate vertical shifts across the page and estimate the overlaps between the blocks. Once the blocks are detected, characters are extracted using CNNs applied to the blocks, as shown in Fig. 4. CNNs have been extensively used for a wide range of image and video processing and visual perception tasks [32], such as object detection, object classification, pattern recognition, and character recognition. Behind the remarkable success of CNNs is their unique capability to extract the underlying nonlinear structures of images by optimizing parameters of multiple layers with the so-called translational invariance [33].

Once individual character recognition has been performed, word recognition and language correctness are improved with linguistic analysis to determine the coherence of text by employing language models. Language models learn statistical features to model the probability distribution over sequences of words [34]. For a sequence formed by words of length n, language models assign a probability $P(w_1, w_2, \ldots, w_n)$ to the whole sequence. Neural language models, or continuous space language models, use continuous representations such as word or character embeddings [35] to compute these probabilities. Usually neural network language models (NNLMs) are constructed and trained as probabilistic classifiers that learn to predict a probability distribution $P(w_t \mid w_{t-k}, \ldots, w_{t-1}) \forall t \in \text{Vocabulary}$. In this approach, the network is trained to predict a probability distribution over the vocabulary, given some linguistic context. The learning process is carried out using standard neural network training algorithms such as stochastic gradient descent and backpropagation [36].

4.3. The corporate relation extraction component

The relation extraction (RE) component processes the unstructured natural text obtained from the previous components and extracts structured relations. This process is performed by following an NLP pipeline composed of tokenization, morphosyntactic and syntactic parsing, named-entity recognition, and information extraction based on finite automata [37,38].

A tokenizer is an NLP resource that breaks unstructured text into chunks of information that can be considered discrete elements. Usually, tokenization is the first step of NLP architectures, and this process outputs occurrences of words from a vocabulary represented as vectors. We follow [39] to tokenize the unstructured text and obtain sequences of tokens. The tokenization rules are based on spaces and specific rules for punctuation marks or special characters in the Spanish vocabulary.

The tokenization process is followed by lemmatization and morphosyntactic parsing, following [39,40]. On the one hand, lemmatization consists of the process of grouping together inflected forms of words so they can be analyzed as single items, identified by the root of the word (the lemma). On the other hand, morphosyntactic and syntactic parsing allow access to information about the linguistic category and dependencies of words (e.g., whether a word is a noun, pronoun, adverb, determiner, preposition, etc.). That is, the parsing process refers to categorizing words in a text in correspondence with a particular part of speech, depending on the definition of the word and its context.

Once the text is normalized, Named-Entity Recognition (NER) is used in order to identify people, organizations, places, and expressions of time and amounts. The area of NER is a subtask of information extraction in the field of NLP that aims to locate and classify named entities mentioned in unstructured text into predefined categories [41]. We follow [42] and build a transition-based named entity recognition system for the identification of the attributes of the relations. Transition-based entity recognizer systems identify non-overlapping labeled spans of words and assume that the most decisive information about entities is found in proximal words. In the literature of NLP, transition-based models with global optimization based on entity accuracy loss functions have been extensively exploited for NER. Transition-based models map structured output construction into an incremental state-transition process, where the score of a whole sequence of transition actions is trained globally in order to resolve structural ambiguities. We define the end-to-end labeled span recognition task as follows: given the processed natural text t and a controlled set of labels L, the task of end-to-end NER is to identify all span mentions in t and link each of the identified mentions with its corresponding concept from L.

[2] https://scrapy.org/.
[3] https://pypi.org/project/pytesserac.
[4] https://github.com/tesseract-ocr/tesseract.

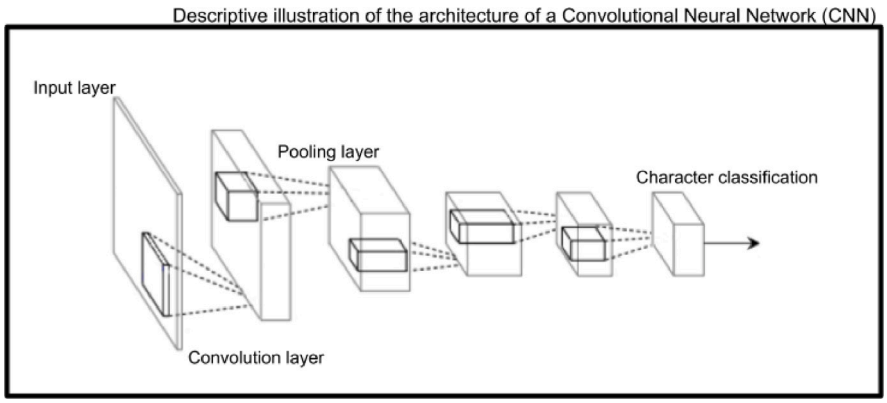


Fig. 4. The CNN architecture for character recognition. CNNs include convolution layers and pooling layers through which the feature maps inherit the dimensionality. A final linear multi-layer perceptron (MLP) classifies the character based on the probability distribution of the whole set of characters [32].

4.3.1. NER component architecture

Following [43], our model applies a shift-reduce parsing approach in which a stack is employed to store partially processed spans. An additional memory buffer is used to store unprocessed tokens, and the learning is framed as predicting the next action given the state of the parser, which can be either: shift, out, complete, reduce, left-reduce, or right-reduce. The *Shift* action moves the first token from the buffer to the stack, the *Out* action pops the first token from the buffer, the *Complete* action pops the top span of the stack, resulting in a new entity mention, the *Reduce* action pops the top two spans from the stack and concatenates them, the *Left-reduce* action performs the same operation as reduce but keeps the second token from the stack, and the *Right-reduce* action performs the same operation as reduce but keeps the first token from the stack.

As stated by [43], when recognizing entity mentions from text, the set of possible actions for the parser is determined by the parser state, and not all types of actions are valid. For instance, the actions *shift* and *out* are only valid when the stack contains a buffered token, as they necessarily require some content to operate on. The parser is trained as described below to learn the most likely actions from the set of valid actions.

Given a sequence of N tokens that must be processed, transition-based neural network NER systems first run a bidirectional LSTM [44] to derive the contextual representation of each token. For each token in the sequence, its representation is denoted as: $\text{Token} = [\overrightarrow{LSTM}(t_0, \ldots, t_i); \overleftarrow{LSTM}(t_i, \ldots, t_{n-1})]$ where t_i indicates the concatenation of the embeddings for the ith token and the semicolon indicates vector concatenation between the forward and backward passing LSTMs. The usage of pretrained contextual word representations have shown a major breakthrough in recent NLP research [45].

Then, the representation of spans is summed with the action moves to obtain the parser states from which the most probable moves are learned. In this way, the NER system concatenates the representation of the top three spans from the stack (s0, s1, s2), as well as the representation of the previous actions, and produces a softmax prediction on the probability distribution of the actions for the next state using a unidirectional one-layer LSTM.

To train the system, we employ the UD Spanish AnCora v2.5 corpus [46], WikiNER [47], and Spacy Lookups data [48]. The transition-based LSTM model obtains SOtA results in identifying entity types such as people, locations, and organizations, offering a global accuracy of 0.89 for the entire set of target categories [40]. Table 1 provides a summary of the datasets and their respective NER labels.

4.3.2. RE component architecture

Applications from the information extraction domain require a complex understanding of the semantic relations between the entities involved. Though NER components are able to automatically label and extract entities with high accuracy, the entire relation extraction process is a more complex task than mere entity recognition. For successful relation extraction, the system not only needs to recognize a relevant piece of text but also understand the semantic properties needed to correctly extract the semantic relation. In the present work, a relation is defined in the form of a tuple, where the full set of 23 relations and their attributes is defined as shown in Tables 2 and 3. Thus, the tuple is composed of a first payload denoting the relation type, and a second payload denoting its attributes. In coordination with the Treasury and Finance Department of the Provincial Council of Gipuzkoa, we configured this predefined set of 23 relevant corporate relations and their respective attributes.

The work of [49] describes different alternatives for relation extraction systems, and the present work exploits the kernel tree approach to extract relations from processed syntactic trees (See Fig. 5). Compared to other approaches, kernel tree models analyze the shallow parse tree or the dependency parse tree built upon the sentence for relation extraction. Thus, once the NER labels are added to the parse tree, the application makes use of finite state automata to identify the keywords predefined in a lookup dictionary that collects the most relevant operations contained in the Official Gazette of the Commercial Registry. Once the root among the parse tree tokens is identified, the finite automata gathers the attributes for the identified corporate relation as shown in Tables 2 and 3. The NER labels of the tokens are used to extract the related attributes from the subsuming nodes of the tree.

As shown by [49], kernel trees are a rather straightforward approach for relation extraction that achieves accurate results, although they also have some limitations: (i) the method is complex to generalize if new relations are to be extracted, and (ii) the method requires a shallow or dependency parse tree to work on and is, consequently, prone to propagate preprocessing pipeline errors.

4.4. The persistence component

The persistence component enables grouping the relations by type and generating a structured knowledge-base containing all corporate relations through a common interface. Thus, relations can be saved in different formats, such as comma-separated values or relational databases, as long as they follow the schema described in Tables 2 and 3.

5. Results and evaluation

To evaluate the performance of the information extraction system, we exploited the date range of the BORME between *2019/01/01* and *2020/04/17*. We performed a massive extraction of all web addresses in the defined time period and attempted to collect all corporate relations within the frame. We manually checked all the extracted relations by

Table 1
Summary statistics for the NER datasets.

Dataset	# Tokens	NER labels					
		Organization	Location	Person	Number	Date	Other
Spanish AnCora v2.5	547,203	X	X	X	X	X	X
WikiNER	254,787,200	X	X	X		X	X
Spacy Lookups	14,995	X	X	X	X		

```
In          < Adposition >     ( Preposition )
compliance  < Noun >     ( Preposition object )
with        < Adposition >     ( preposition )
the         < Determiner >     ( Determiner )
provisions  < Noun >     ( Preposition object )
of          < Adposition >     ( Preposition )
article     < Noun >     ( Preposition object )
43          < Numeral >     ( Pronominal quantifier )
of          < Adposition >     ( preposition )
Law         < Proper noun >     ( Preposition object )
...
as          < Subordinating conjunction >     ( Preposition )
side        < Noun >     ( Compound )
partner     < Noun >     ( Preposition object )
of          < Adposition >     ( Preposition )
the         < Determiner >     ( Determiner )
company     < Noun >     ( Preposition object )
Anonymized  < Proper noun >     ( Preposition object )
it          < Pronoun >     ( Nominal subject (passive) )
was         < Auxiliary >     ( Auxiliar (passive) )
agreed      < Verb >     ( Root )
to          < Particle >     ( Auxiliar )
approve     < Verb >     ( Open clausal complement )
...
```

Fig. 5. Result of tokenizing and parsing the text contained in Fig. 2. Words are listed along their part-of-speech tag and their corresponding universal sentence dependencies.

Table 2
Structure of the 16 corporate relations extracted from the first section of the BORME. Capital c stands for Corporation.

Corporate relation type	Attributes of the relation
Expansion of the corporate purpose	C., New corporate purpose, Date
Change of C. name	C. before change, C. after change, Date
Change of registered office	C., New registered office, Location, Date
Unique partner identity change	C., New unique partner, Date
Global transfer of assets and liabilities	Assigner C., Assignee C., Date
Incorporation of C.	C., Corporate purpose, Address, Start operations, Location, Capital, Date
Payment of capital calls	C., Payment, Date
State of insolvency	C., Date
Split-off	Divided C., Receiving C., Date
Extinction	C., Date
Errata sheet	C., Text, Date
Mergers by absorption	Acquiring C., Acquired C., Date
Non-approval of accounts	C., Fiscal year, Date
Appointment	C., Appointed person, Position, Date
Reelection	C., Reelected person or C., Position, Date
Revocation	C., Revoked person, Position, Date

Table 3
Structure of the 7 corporate relations extracted from the second section of the BORME. Capital c stands for Corporation.

Corporate relation type	Attributes of the relation
Absorption	Absorbing C., Absorbed C., Date
Capital increase	C., Capital, Date
Transfer	Assigner C., Assignee C. , Date
Dissolution	C., Date
Liquidation	C., Date
Reactivation	C., Date
Capital reduction	C., Capital, Date

comparing both the correctness of the relations and the correctness of the attributes against the explicit contents of the BORME.

Tables 4 and 5 show the type and number of relations extracted from the first and second sections of the BORME for the province of Gipuzkoa in the specified date range. Specifically, 1,056 relations were extracted from the first section of the BORME and 268 relations from the second section.

Table 6 reports the results of the evaluation of the performance of the information extraction system. This table shows the detailed results for each of the types of relations extracted from the first and second sections of the BORME, indicating the total number of relations extracted (Rels), the number of relations extracted correctly (True Positive or TP), the number of incorrectly extracted relations (False Positive or FP) and the precision of the extraction system (P). This precision is calculated according to Eq. (1) which follows SOtA in the field of information extraction [50]. Moreover, relations are considered to be extracted incorrectly if the relation type or the attributes of the relation are not correct.

$$\text{P} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Macro P} = \frac{\sum_{relations} P_{relation}}{Number\ relations} \tag{2}$$

$$\text{Micro P} = \sum_{relations} \frac{Instances_{relation}}{Total\ Instances} \times P_{relation} \tag{3}$$

Table 4
Number and proportion of relations extracted from the first section of the BORME between *2019/01/01* and *2020/04/17* for the province of Gipuzkoa.

Relation type	# Rels	Per-unit
Expansion of the corporate purpose	70	0.07
Change of corporation name	71	0.07
Change of registered office	121	0.11
Unique partner identity change	72	0.07
Global transfer of assets and liabilities	86	0.08
Incorporation of corporation	94	0.09
Payment of capital calls	9	0.01
State of insolvency	67	0.06
Split-off	45	0.04
Extinction	80	0.08
Errata sheet	31	0.03
Mergers by absorption	50	0.05
Non-approval of accounts	63	0.06
Appointment	115	0.011
Reelection	18	0.02
Revocation	64	0.06
Total	1,056	1

Table 5
Number and proportion of relations extracted from the second section of the BORME between *2019/01/01* and *2020/04/17* for the province of Gipuzkoa.

Relation type	# Rels	Per-unit
Absorption	69	0.26
Capital increase	62	0.23
Transfer	31	0.12
Dissolution	9	0.03
Liquidation	5	0.02
Reactivation	24	0.09
Capital reduction	68	0.25
Total	268	1

Table 6
Precision of the relation extraction system for the first (above double line) and second sections (below double line) of the BORME between *2019/01/01* and *2020/04/17* for the province of Gipuzkoa.

Relation type	# Rels	# TP	# FP	P
Expansion of the corporate purpose	70	70	0	1
Change of corporation name	71	71	0	1
Change of registered office	121	121	0	1
Unique partner identity change	72	72	0	1
Global transfer of assets and liabilities	86	86	0	1
Incorporation of corporation	94	94	0	1
Payment of capital calls	9	9	0	1
State of insolvency	67	65	2	0.971
Split-off	45	44	1	0.977
Extinction	80	80	0	1
Errata sheet	31	31	0	1
Mergers by absorption	50	50	0	1
Non-approval of accounts	63	63	0	1
Appointment	115	115	0	1
Reelection	18	18	0	1
Revocation	64	64	0	1
Absorption	69	65	4	0.942
Capital increase	62	53	9	0.854
Transfer	31	29	2	0.935
Dissolution	9	9	0	1
Liquidation	5	5	0	1
Reactivation	24	24	0	1
Capital reduction	68	60	8	0.882

Overall results indicate that the precision of the system is high for all the relations and attributes extracted, as the macro-precision is 0.983 (Eq. (2)) and the micro-precision is 0.99 (Eq. (3)). The relations that have shown greater difficulty for extraction compound to split-offs (0.977), state of insolvencies (0.971), absorptions (0.942), transfers (0.935); and especially capital increases and reductions, with a 0.854 and 0.882 precision respectively.

We performed an in-depth error analysis of the incorrectly extracted relations and observed that most of the cases correspond to the incorrect identification of the 'amount' attribute. We noted that in this particular kind of incorrect extraction, the system fails to correctly analyze quantity data, as the attribute is given in non-standard ways, combining numerical and textual descriptions (e.g., '3 thousand' instead of '3000' or 'three thousand'). Other errors relate to corporation, beneficiary, or person name identification issues inherited from NLP resources (e.g., company or person name partly identified). Although some relations, such as capital reductions and increases, show lower precision, the system demonstrates a high overall precision rate, indicating that an effective extraction system can be developed using NLP and CV techniques to generate a knowledge-base composed of corporate relations.

6. Conclusions and future work

Tax fraud is a criminal activity that causes significant income loss for public institutions. The clandestine nature of tax evasion makes it difficult to quantify this loss, although the most conservative estimates place it at around 5% of world GDP. The fight against tax fraud requires technological tools that enable the investigation and validation of data from multiple sources, such as tax agencies and unstructured knowledge-bases from external sources, to detect inconsistencies or patterns that can help identify possible sources of tax fraud.

Our work contributes to the fight against tax fraud by introducing an automated corporate relation extraction system capable of processing large volumes of unstructured data from official sources. The construction of such a knowledge-base responds to a need raised by the Treasury and Finance Department of the Provincial Council of Gipuzkoa which required a procedure to cross-reference existing tax information from the agency with data available in the Commercial Registry in order to identify possible inconsistencies. To this end, we construct and evaluate a relation extraction system capable of generating 23 structured corporate relations from unstructured sources, such as the BORME. The tool is composed of NLP and CV technologies that enable fast processing of large volumes of data, helping to automate the task for expert human auditors. The system has demonstrated high precision across all extracted relations underscoring its potential as a valuable resource in automating tax fraud detection efforts.

The extraction system is based on four components that run in a waterfall model and generate structured corporate relations from unstructured web sources. We manually evaluate the system by verifying the extracted relations with regard to the original content in the BORME and demonstrate that our system attains high precision for all 23 relations and their attributes, indicating that the relation extraction system is a valuable resource in the fight against tax fraud.

For future work, it would be appropriate to incorporate new subsystems into the current extraction tool to increase its coverage and precision, including additional corporate relationships and attributes. Likewise, it would also be worthwhile to expand the geographical scope and the evaluation period, and to design automated procedures to speed up the manual evaluation, since the human effort required for this task is substantial. An exploratory analysis of how many relation mentions

were missed by the trigger-phrase-based RE approach, which requires a large amount of manual effort, is also a line of future work.

CRediT authorship contribution statement

Inigo Lopez-Gazpio: Conceptualization, Data curation, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Laura Baselga-Pascual:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Aitor Garmendia-Lazcano:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was conducted as part of the Projects PID2022-136818NB-100 and PID2021-122133NB-I00 financed by MCIN/AEI/10.13039/501100011033/FEDER, EU. We also gratefully acknowledge financial support from the Basque Government Department of Education (IT1497-22 y IT1570-22).

Data availability

The data contained in the BORME used for the present research is publicly available at: https://www.boe.es/diario_borme/.

References

[1] G.S. Becker, Crime and punishment: An economic approach, J. Polit. Econ. 76 (1968) 169–217.
[2] H. Thakkar, S. Datta, P. Bhadra, H. Barot, M. Purohit, S. Dabhade, A bibliometric analysis of forensic accounting research: Unveiling its impact on tax fraud detection in SAARC countries, J. Inform. Educ. Res. 4 (2) (2024).
[3] J.S. Henry, The price of offshore revisited, Tax Justice Netw. 22 (2012) 57–168.
[4] R. Murphy, T. Riley, The Cost of Tax Abuse: A Briefing Paper on the Cost of Tax Evasion Worldwide, The Tax Justice Network, Institute for Economic Affairs, London, 2011.
[5] F. Bonchi, F. Giannotti, G. Mainetto, D. Pedreschi, Using data mining techniques in fiscal fraud detection, in: International Conference on Data Warehousing and Knowledge Discovery, Springer, 1999, pp. 369–376.
[6] X.D. Zhang, Machine learning, in: A Matrix Algebra Approach To Artificial Intelligence, Springer, 2020, pp. 223–440.
[7] E. Tenidou, S. Valsamidis, I. Petasakis, A. Mandilas, Elenxis, an effective tool for the war against tax avoidance and evasion, Procedia Econ. Financ. 33 (2015) 303–312.
[8] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, Prentice Hall, 2002.
[9] K. Malaszczyk, B.M. Purcell, Big data analytics in tax fraud detection, Northeast. Assoc. Bus. Econ. Technol. (2017) pp. 233.
[10] A. Nawawi, A. Salin, S.A. Puteh, Capital statement analysis as a tool to detect tax evasion, Int. J. Law Manag. (2018).
[11] Q. Zheng, Y. Xu, H. Liu, B. Shi, J. Wang, B. Dong, A survey of tax risk detection using data mining techniques, Engineering 34 (2024) 43–59.
[12] N. Alsadhan, A multi-module machine learning approach to detect tax fraud, Comput. Syst. Sci. Eng. 46 (1) (2023) 241–253.
[13] L.N. Lata, I.A. Koushika, S.S. Hasan, A comprehensive survey of fraud detection techniques, Int. J. Appl. Inf. Syst. 10 (2) (2015) 26–32.
[14] M. Van Banerveld, N.A. Le-Khac, M.T. Kechadi, Performance evaluation of a natural language processing approach applied in white collar crime investigation, in: International Conference on Future Data and Security Engineering, Springer, 2014, pp. 29–43.
[15] Y. Kou, C.T. Lu, S. Sirwongwattana, Y.P. Huang, Survey of fraud detection techniques, in: IEEE International Conference on Networking, Sensing and Control, 2004, Vol. 2, IEEE, 2004, pp. 749–754.
[16] R. Smith, An overview of the tesseract OCR engine, in: Ninth International Conference on Document Analysis and Recognition, Vol. 2, ICDAR 2007, IEEE, 2007, pp. 629–633.
[17] S. Castilho, J. Moorkens, F. Gaspari, I. Calixto, J. Tinsley, A. Way, Is neural machine translation the new state of the art? Prague Bull. Math. Linguist. 108 (1) (2017) 109–120.
[18] N.I. Altmami, M.E.B. Menai, Automatic summarization of scientific articles: A survey, J. King Saud Univ.- Comput. Inf. Sci. (2020).
[19] A. Doan, R. Ramakrishnan, S. Vaithyanathan, Managing information extraction: state of the art and research directions, in: Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data, 2006, pp. 799–800.
[20] O. Feyisetan, M. Luczak-Rösch, E. Simperl, R. Tinati, N. Shadbolt, Towards hybrid NER: a study of content and crowdsourcing-related performance factors, in: European Semantic Web Conference, Springer, 2015, pp. 525–540.
[21] A. Auger, C. Barrière, Pattern-based approaches to semantic relation extraction: A state-of-the-art, Terminology 14 (1) (2008) 1.
[22] S. Basta, F. Fassetti, M. Guarascio, G. Manco, F. Giannotti, D. Pedreschi, L. Spinsanti, G. Papi, S. Pisani, High quality true-positive prediction for fiscal fraud detection, in: 2009 IEEE International Conference on Data Mining Workshops, IEEE, 2009, pp. 7–12.
[23] S. Smith, M.M. Keen, VAT Fraud and Evasion: What Do We Know, and What Can Be Done?, No. 7–31, International Monetary Fund, 2007.
[24] S. Abney, Partial parsing via finite-state cascades, Nat. Lang. Eng. 2 (4) (1996) 337–344.
[25] C. Manning, H. Schutze, Foundations of Statistical Natural Language Processing, MIT Press, 1999.
[26] J. Memon, M. Sami, R.A. Khan, M. Uddin, Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR), IEEE Access 8 (2020) 142642–142668.
[27] R.C. Martin, Agile Software Development: Principles, Patterns, and Practices, Prentice Hall, 2002.
[28] R. Chen, Z. Wang, H. Su, S. Xie, Z. Wang, Parallel xpath query based on cost optimization, J. Supercomput. (2021) 1–30.
[29] R. Smith, A simple and efficient skew detection algorithm via text row accumulation, in: Proceedings of 3rd International Conference on Document Analysis and Recognition, Vol. 2, IEEE, 1995, pp. 1145–1148.
[30] W. Cho, J. Kwon, S. Kwon, J. Yoo, A comparative study on OCR using super-resolution for small fonts, Int. J. Adv. Smart Converg. 8 (3) (2019) 95–101.
[31] S. Dhiman, A. Singh, Tesseract vs gocr a comparative study, Int. J. Recent. Technol. Eng. 2 (4) (2013) 80.
[32] H. Lee, H. Kwon, Going deeper with contextual CNN for hyperspectral image classification, IEEE Trans. Image Process. 26 (10) (2017) 4843–4855.
[33] T. Wiatowski, H. Bölcskei, A mathematical theory of deep convolutional neural networks for feature extraction, IEEE Trans. Inform. Theory 64 (3) (2017) 1845–1866.
[34] L.R. Bahl, P.F. Brown, P.V. de Souza, R.L. Mercer, A tree-based statistical language model for natural language speech recognition, IEEE Trans. Acoust. Speech Signal Process. 37 (7) (1989) 1001–1008.
[35] X. Chen, L. Xu, Z. Liu, M. Sun, H. Luan, Joint learning of character and word embeddings, in: Twenty-Fourth International Joint Conference on Artificial Intelligence, 2015, pp. 115–123.
[36] L. Bottou, Stochastic gradient descent tricks, in: Neural Networks: Tricks of the Trade, Springer, 2012, pp. 421–436.
[37] R. Agerri, J. Bermudez, G. Rigau, IXA pipeline: Efficient and ready to use multilingual NLP tools, in: LREC, Vol. 2014, 2014, pp. 3823–3828.
[38] X. Zhao, Y. Deng, M. Yang, L. Wang, R. Zhang, H. Cheng, W. Lam, Y. Shen, R. Xu, A comprehensive survey on deep learning for relation extraction: Recent advances and new frontiers, 2023, arXiv preprint arXiv:2306.02051.
[39] N. Hardeniya, J. Perkins, D. Chopra, N. Joshi, I. Mathur, Natural Language Processing: Python and NLTK, Packt Publishing Ltd, 2016.
[40] X. Schmitt, S. Kubler, J. Robert, M. Papadakis, Y. LeTraon, A replicable comparison study of NER software: StanfordNLP, NLTK, OpenNLP, SpaCy, Gate, in: 2019 Sixth International Conference on Social Networks Analysis, Management and Security, SNAMS, IEEE, 2019, pp. 338–343.
[41] T.M. Georgescu, B. Iancu, A. Zamfiroiu, M. Doinea, C.E. Boja, C. Cartas, A survey on named entity recognition solutions applied for cybersecurity-related text processing, in: Proceedings of Fifth International Congress on Information and Communication Technology, Springer, 2021, pp. 316–325.
[42] Y. Lou, Y. Zhang, T. Qian, F. Li, S. Xiong, D. Ji, A transition-based joint model for disease named entity recognition and normalization, Bioinformatics 33 (15) (2017) 2363–2371.
[43] X. Dai, S. Karimi, B. Hachey, C. Paris, An effective transition-based model for discontinuous NER, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 5860–5870, http://dx.doi.org/10.18653/v1/2020.acl-main.520, Online. URL https://www.aclweb.org/anthology/2020.acl-main.520.

[44] A. Graves, Long short-term memory, in: Supervised Sequence Labelling with Recurrent Neural Networks, Springer, 2012, pp. 37–45.
[45] J. Goikoetxea, E. Agirre, A. Soroa, Single or multiple? combining word representations independently learned from text and wordnet, in: Thirtieth AAAI Conference on Artificial Intelligence, 2016.
[46] H.M. Alonso, D. Zeman, Universal dependencies for the AnCora treebanks, Proces. Del Leng. Nat. 57 (2016) 91–98.
[47] J. Nothman, N. Ringland, W. Radford, T. Murphy, J.R. Curran, Learning multilingual named entity recognition from wikipedia, Artificial Intelligence 194 (2013) 151–175.
[48] A. Explosion, spaCy-industrial-strength natural language processing in python, 2017, URL https://spacy.io.
[49] N. Bach, S. Badaskar, A review of relation extraction, Lit. Rev. Lang. Stat. II 2 (2007) 1–15.
[50] J. Piskorski, R. Yangarber, Information extraction: Past, present and future, in: Multi-Source, Multilingual Information Extraction and Summarization, Springer, 2013, pp. 23–49.