24th International Conference on Business Information Systems (BIS 2021)

Big Data

https://doi.org/10.52825/bis.v1i.44

Published: 02 July 2021

Developing a legal form classification and extraction approach for company entity matching

Benchmark of rule-based and machine learning approaches

Felix Kruse [1][https://orcid.org/0000-0001-8033-0840], Jan-Philipp Awick [1], Jorge Marx Gómez [1], and Peter Loos [2]

[1] University of Oldenburg

[2] DFKI & Saarland University

Abstract. This paper explores the data integration process step record linkage. Thereby we focus on the entity company. For the integration of company data, the company name is a crucial attribute, which often includes the legal form. This legal form is not concise and consistent represented among different data sources, which leads to considerable data quality problems for the further process steps in record linkage. To solve these problems, we classify and ex-tract the legal form from the attribute company name. For this purpose, we iteratively developed four different approaches and compared them in a benchmark. The best approach is a hybrid approach combining a rule set and a supervised machine learning model. With our developed hybrid approach, any company data sets from research or business can be processed. Thus, the data quality for subsequent data processing steps such as record linkage can be improved. Furthermore, our approach can be adapted to solve the same data quality problems in other attributes.

Keywords: Record Linkage, Company Entity Matching, Data Integration, Data Quality, Data Preparation.

Motivation and problem statement

Companies try to integrate data in their decision-making processes in the most efficient way to achieve corporate added value. The cyclical process of the information value chain describes this approach of the companies. First, data is transferred into information and then into knowledge. This knowledge is used in decision-making processes and subsequent actions to generate added value for the company [1]. The information advantage becomes a crucial part of a company's economic success. Heinrich and Stühler [2] showed that companies that integrate relevant data directly into their decision-making processes are more competitive [2]. For example, data about competitors, suppliers, or corporate customers may contain such company and competition relevant information. However, this information is often hidden in multiple external and internal data sources [3–5]. In many cases, only the combination of external and internal data sources leads to interesting, novel, valuable, and unexpected insights that provide a competitive advantage [4–6].

The data integration goal is to provide unified access to these external and internal data sources [6]. The data integration process consists of the process steps (1) schema matching, (2) record linkage (RL), and (3) data fusion to achieve this goal. The schema matching step serves to identify the attributes that have the same meaning [6]. The RL step matches data records from different data sources that refer to the same real-world entity such as companies, products, or persons [7, p. 3-4]. The data fusion step determines the valid values

of the respective attributes of the matched record [6]. While the data integration goal is easy to formulate it is still hard to achieve [6]. RL is a crucial task in the data integration process and has become a sub-discipline of data science due to its complexity and similarity to classical data science tasks [8–10]. The complexity is caused when unique identification numbers among the data sources are missing and other existing attributes are very heterogeneous. If no identification number exists, RL must be performed using additional attributes in the data sources. For competitors, suppliers, or enterprise customers, these are the company name, the address, or the company description [11]. For RL, market participants such as competitors, suppliers, or corporate customers represent the real-world entity company. RL is still very messy in practice since there are many data sources containing different real-world entities, this leads to many RL scenarios with several challenges [8]. Köpcke *et al.* [12] try to reduce the multitude of RL scenarios' complexity by focusing on the real-world entity product [12]. Our RL research focuses on the real-world entity company. We define this as company entity matching. We have identified several RL challenges within the existing attributes company name, address data and company description [13]. We identified these challenges through our data-driven inductive research method [14]. This method describes our approach to analysing our eleven existing data sources (see table 1) and integrating various of them through a RL process to find general RL challenges for the real-world entity company.

Table 1: Company data sources for inductive data-driven research

Data source	Source
Handelsregister	https://offeneregister.de/
OpenCorporates	https://opencorporates.com/
Crunchbase ODM	https://data.crunchbase.com/docs/open-data-map
Crunchbase Snapshot	https://data.crunchbase.com/docs/2013-snapshot
GLEIF	https://www.gleif.org/en
USPTO	https://developer.uspto.gov/
Wikidata	https://www.wikidata.org/
Uscompanylist - Company	https://www.uscompanieslist.com/
Uscompanylist - Business	https://www.uscompanieslist.com/
AlphaVantage	https://www.alphavantage.co/
Owler	https://corp.owler.com/

One of the most relevant attributes in company entity matching is the company name [15, 16] which we will focus on in this paper. The legal form of a company is also an important attribute, as it is discriminatory when comparing companies [15]. In our eleven data sources (see table 1), the company legal form is always contained in the company name attribute, as the nine examples in table 2 show. The company name contains the company's legal form and thus is not atomic. This leads to the problem that the attribute legal form cannot be directly analysed without further data preparation efforts. Wang and Strong [17] formalize this as a not concise representation of the data and thus as a data quality problem. Besides, the company legal form is often represented inconsistently. Table 2 shows nine different representations for the German company legal form "GmbH". The nine records show punctuation problems, upper- and lower-case problems, abbreviation problems, and umlaut problems. The legal form is not always at the end of a company name (ID 5), and the legal form tokens can be separated by tokens of the company name (ID 9). The consistent representation is also defined as a data quality dimension by Wang and Strong [17]. The inconsistent representation of the legal form leads to the problem that, for example, the analysis of a particular legal form like the "GmbH" requires much effort. In addition, company names can be represented differently in various databases due to the inconsistent representation of the legal form. This makes the company entity matching more difficult. The two data quality problems of concise and consistent representation of the legal form are even more complicated as various legal forms exist for each country in the world.

Table 2: Different representation of the legal form "GmbH" in the company name

ID	Company_name
1	Selbstfahrer Union **G.m.b.H.**
2	GIANT Weilerswist g21 **GmbH**
3	FABIUS Vermietungs**gesellschaft mbH**
4	Infrastrukturentwicklungs**gesellschaft** Hilden **mbH**
5	ITM & C **GmbH** International Trade Marketing & Consulting
6	FHS Gabelstapler **Gesellschaft mit beschränkter Haftung**
7	bunse aufzuege **gesellschaft mit beschraenkter haftung**
8	alint 458 grundstueckverwaltung **gesellschaft m.b.h.**
9	**gesellschaft** zur verwertung von leistungsschutzrechten **mit beschraenkter haftung** gvl

Figure 1 shows the company entity matching challenges when the legal form is included in the company name, which we identified through our inductive data-driven experiments, and the matching when the company name and legal form are split into different attributes. First, we discuss the matching problems when the company name and legal form are within the same attribute. There are two data sources, A and B, with two companies that differ only in their legal form. As a human being, it is obvious that the tuples with the ID's C100 and 2 and C101 and 1 belong to the same entity. The classic string similarity measures such as normalized Levenshtein, Jarowinkler, Jaccard, or Soft TF/IDF [18] do not provide exact results to determine match and no-match tuples. The highest values of the normalized Levenshtein distance would classify the two non-match tuples as matches. The highest values of the Jarowinkler Distance would combine a match (ID C101 and 1) and a non-match (ID C100 and 1). The Jaccard distance does not distinguish the tuples. The Soft TF/IDF distance does not distinguish the tuples for the company with the ID C100. For the company with the ID C101, the higher Soft TF/IDF would be the match. With this small example, the problem for company entity matching with the legal form within the company name attribute is shown. However, with the second example shown in figure 1 where company name and legal form are split into two separate attributes, all string similarity measures show a similarity of 100% for the cleaned name, but the legal form is only the same for the matches. This allows the correct tuples to be selected as matches. This shows that a data preparation approach is needed to split the company name into the attributes company name without legal form (cleaned name) and company legal form (legal form) to achieve our goal.

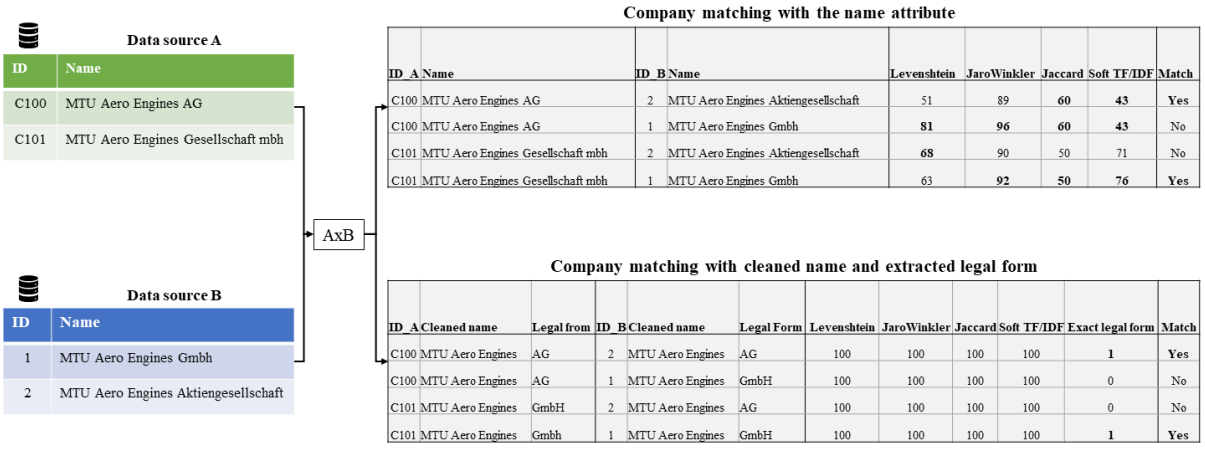

Figure 1: Example of the legal form problem with company entity matching

Our paper therefore aims to develop an approach that classifies and extracts the company name's legal form to improve the data quality and support further data processing steps such as the RL. Note that we focus on the German legal forms as a starting point for our research. Based on this introduction and problem statement, we address the research question:

"Which approach is appropriate to classify and extract the legal form in the company name?" To answer the research question, we follow the inductive data-driven research approach according to [14]. This approach seems to us to be most suitable for data science research, because Maass *et al.* [19] defines data-driven research as "an exploratory approach that analyses data to extract scientifically interesting insights (e.g., patterns) by applying analytical techniques and modes of reasoning". To carry out the research approach, we iteratively identified, implemented, and evaluated potential approaches and analysed the data to extract scientifically insights about the best performing approach. We have tried to improve the approaches or identify new approaches until the results were acceptable. We present the results of the developed approaches in a summarising benchmark.

The paper is structured as follows. Section 2 describes the related work. In section 3, we present our four identified and implemented approaches for the benchmark. In section 4, we describe and analyse the results of our conducted benchmark. In section 5, we present the theoretical and practical implications and the limitations of our paper. The paper ends in section 6 with a conclusion and outlook.

Related Work

The process steps (1) data preparation, (2) blocking, (3) record pair comparison, (4) classification, and (5) evaluation perform RL [7, p. 24, 20]. The literature review by Kruse *et al.* [20] shows that the focus of current research in the field of RL is primarily on the process step classification and the entire RL process for a given data source pair.
In general RL research, there are RL approaches that achieve high F1-scores [21–23] on existing RL datasets provided mainly by the Magellan project [8, 24]. Nevertheless, we cannot compare our research with these results because the used datasets do not consider the RL challenge of company legal form that we identified. Most of the datasets are used to link the real-world entities product or person in which the company legal form does not exist. Only one dataset is used to link the real-world entity company[1]. Mudgal *et al.* [22] classify the dataset as a textual dataset because the dataset has only a unstructured company description as an attribute. This attribute does not have the RL problem we identified with the company legal form, as we narrow this problem down to the structured attribute company name.

Since no benchmark dataset exists to test our approach against other RL processes, we initially classify the work in the research area of company entity matching and data preparation in RL. This paper deals with the process step data preparation, which has been little researched in the context of company entity matching. The papers identified in the areas of company entity matching and data preparation for RL are presented below.

Company Entity Matching

We have identified the papers of Schild and Schultz [15], Cuffe and Goldschlag [25], and Gschwind *et al.* [16] as research papers that focus specifically on company entity matching. Schild and Schultz [15] present in their paper a self-developed RL process to integrate different data sources containing companies for research purposes of the Deutsche Bundesbank. In the paper, seven data sources are used. Two data sources are provided by external data providers Bureau van Dijk and Hoppenstedt/Bisnode. Five are internal data sources of the Bundesbank. The attributes company name, legal form, postal code, city, and street were used for RL. Schild and Schultz [15] describe the company name as the most important attribute to distinguish company entities. The company name's distinctiveness can be enhanced by geographical additions or the legal form in the company name. For Schild and Schultz [15], the most important attribute for comparing companies is their legal form. They have developed a set of rules consisting of regular expressions to classify the legal form. The set of rules classifies only the german legal forms. Schild and Schultz (2017) research results show the importance of company matching for subsequent analytical use

[1] https://github.com/anhaidgroup/deepmatcher/blob/master/Datasets.md#company

cases and the impact of the legal form. In the paper, a very static set of rules for the classification of legal forms was implemented. The approach cannot extract the legal form from the company name. Our approach aims to classify the legal form and to generate a company name without the legal form.

Cuffe and Goldschlag [25] address the problem that there are many individual RL methods and approaches and try to consolidate them in a framework called MAMBA (Multiple Algorithm Matching for Better Analytics). They focus on the entity company and use Census microdata. MAMBA's focus is on applying different string similarity measures in machine learning methods to improve RL results. Cuffe and Goldschlag [25] do not focus on data preparation and thus do not deal with the classification and extraction of the company name's legal form.

Gschwind *et al.* [16] focus on company entity matching to integrate data sources needed for further processing, such as data analytics. The attributes company name, location, and industry are used. The result of the paper is a practical end-to-end system. They used a rule-based approach for the RL process step (4) classification. They developed a machine learning (ML) method for the data preparation process step, which generates a short company name from the company name. For example, the ML method extracts the short company name "Aston Martin" from the company name "Aston Martin Lagonda Limited". The authors defined this problem as a sequence labeling task and trained a conditional random field algorithm to identify and extract the company short name. Gschwind *et al.* [16] do not deal with the company legal form in their ML method to extract the short company name. The company's legal form rarely appears in the company's short name in their case. Their first approach to the RL process deleted the legal form like "inc." or "ltd. However, they observed that some companies only differ in their legal form. As a solution, they have given less weight to the legal form in their scoring process. The paper does not describe how the legal form is identified to give it a lower weight. Neither is the legal form extracted to use it as an additional attribute in the RL process for the record pair comparison.

Data Preparation for Record Linkage

Randall *et al.* [26] and Koumarelas *et al.* [27] focus on the RL process step data preparation. Randall *et al.* [26] examine the effect of data preparation on RL quality. Based on a review by Linkage Software, they identified a set of different data preparation procedures. They applied them to a synthetic dataset and a real administrative dataset to compare RL quality with and without data preparation. The results show that data preparation has little impact on RL quality. The paper does not consider the entity company. The data preparation methods used were very general. The authors themselves say that additional data sets need to be evaluated to make a final statement about the negative or positive impact of data preparation on RL quality. Randall *et al.* [26] call in their outlook that further research should be conducted in more specialized processing of name and address attributes.

Koumarelas *et al.* [27] present a process to select the data preparation methods that improve RL quality the most. The process should contribute to the comparability of future evaluation results in RL research since the description of data preparation is not yet sufficient for this purpose, according to Koumarelas *et al.* [27]. The data preparation procedures considered by Koumarelas *et al.* [27] do not refer to the company name or legal form. General data preparation methods such as "split attributes", "remove special characters" or address related data methods are considered. There is no focus on the entity company for the RL process.

Approaches to classify and extract the legal form

The related work shows no data preparation approach to classify the legal form of a company and extract it from the company name. Furthermore, it shows that specific data preparation can lead to a general increase in data quality and an increase in RL quality. We

have adapted and further developed the approaches Bundesbank and Cleanco and developed two completely new approaches, we called them Deep Learning approach and Hybrid approach, and benchmarked them to classify and extract German company legal forms from company names. In the following, these four developed approaches are described.

Adapted rule-based approach from Bundesbank

The Bundesbank approach is a rule-based approach to classify the company legal form based on [15]. Despite the restriction that the approach only classifies the legal form and does not extract it, it should be included in the benchmark. Schild and Schultz [15] have described the regular expressions in the appendix of their paper and the set of rules to combine certain regular expressions to determine the legal form. The syntax of the regular expressions described in the paper is PERL. They implemented the following German legal forms "GmbH", "AG", "SE", "KG", "OHG", "UG", "GbR", "e.V.", "e.G.", "KGaA", "VVaG.", "GmbH & Co. KG", "GmbH & Co. KGaA", "GmbH & Co. OHG" and "SE". We have implemented the regular expressions and the set of rules in Python. Figure 2 shows the regular expression in Python syntax for the legal form "GmbH". This example illustrates how complicated the regular expressions are to read and extend. For example, regular expressions were developed to classify the legal form "GmbH" and "KG" and the addition "& Co.". If the three regular expressions are found together in a company name, the set of rules classifies a "GmbH & Co. KG". Due to the high modeling effort required to extend the Bundesbank approach e.g. with the legal form extraction function and to add more legal forms, we focused on improving the necessary manual effort in the next approaches.

```
patternGMBH = "(GMBH)|(GmbH)|(G|g)?(E|e)?(S|s)?\.?(ELL|ell)?\.?(SCH|sch)?\.?(AFT|aft)?
              ?(M|m)\.?(IT|it)? ?(B|b)(ESCHR|eschr)?\.?(Ä|ä)?(AE?|ae?)?(NKTER|nkter)?
              ?(H|h)( |,|\.|$|AFTUNG|aftung)"
```

Figure 2: Regular expression for the classification of legal form

Adapted rule-based approach by Cleanco

The Cleanco approach is based on the Github project Cleanco. We identified this project through an internet search for approaches to classify and extract legal forms. Cleanco is a Python-based package that identifies the legal form, removes it, and returns a cleaned company name. Cleanco is based on a rule-based approach. The Bundesbank approach presented in section 3.1 could only classify German legal forms. The Cleanco set of rules contains legal forms from 66 different countries. In our benchmark. Since Cleanco contains German legal forms it was considered in the benchmark. By default, Cleanco has only implemented the German legal forms 'gmbh & co. kg', 'gmbh & co. kg', 'e.g.', 'e.v.', 'gbr', 'ohg', 'partg', 'kgaa', 'gmbh', 'g.m.b.h.' and 'ag'. Besides, Cleanco standardizes all legal forms from different countries to the English legal forms. For example, a "GmbH" is classified as its English equivalent "Limited".

To enable a benchmark with the other approaches, we have adapted and expanded the German legal forms in the Cleanco rules. The legal form "gmbh & co. kg" is implemented in the Cleanco Standard package but is missing in our Cleanco rule set. Since the current implementation of Cleanco cannot classify legal forms consisting of several tokens like "gmbh & co. kg" caused by the technical implementation. For this reason, we had to remove all legal forms that consist of several tokens. Due to the high modelling effort required to remove the technical restriction of classifying legal forms that only consist of one token ("GmbH" works but "GmbH & Co. KG" does not) we have focused on another approach.

Deep Learning (DL) Approach

To implement the deep learning (DL) approach we define the legal form classification and extraction as a sequence labeling problem, such as part-of-speech tagging or named entity recognition [28]. A Sequence Labeling Problem exists if a label from a defined label set is

assigned to each token of a sequence [28]. In our case, the sequence of tokens is the company name. These tokens are to be labeled whether they belong to a specific legal form or not. For sequence labeling, a tagging scheme has to be chosen [29]. In our case, we have chosen the conventional BIO tagging scheme [29, 30]. A starting tag (B), an inner tag (I), and an outside tag (O) is defined. Each of the 27 legal forms (see table 5) is provided with a beginning tag (B-legal form) and an inner tag (I-legal form). All tokens which do not belong to a legal form are assigned the tag (O). An example is shown in table 3.

Table 3: Example of the Sequence Labeling Schema

Name	kuhn	gmbh	facilities	management		
Label	O	B-Gmbh	O	O		

Name	agl	maschinenbau	gesellschaft	mit	beschraenkter	haftung
Label	O	O	B-Gmbh	I-Gmbh	I-Gmbh	I-Gmbh

We created a sample of 10,000 company names based on the GLEIF, Crunchbase ODM, and OpenCorporates databases (see Table 1) to create the training data set. We filtered the databases for German companies. We used the labeling tool *doccano* to label the 10,000 company names with our BIO tagging scheme [31]. We have identified other legal forms such as "Stiftung" or "EK" during the labeling process that are not implemented in the previous approaches Bundesbank and Cleanco. We have created a balanced labelled data set with 18.300 company names. The neural network architectures is a classical bi-directional LSTM (BI-LSTM), often used for sequence labeling problems [28, 30]. The labeled company data set was divided into 80% training data and 20% test data. The BI-LSTM with the best parameter settings achieved an F1 score of about 99.2% on the test data. The DL approach delivers good results but has problems with some legal forms, such as the "gGmbH" and the "PartG", which are often wrongly classified as "GmbH". In addition, sequence labeling according to the BIO tagging scheme requires a high manual effort, as each token in the company name has to be tagged with a label. In order to solve the problems mentioned and expand other legal forms in the future, a high manual labeling effort is necessary. For these reasons, we have developed another approach that should achieve the same or better results, involves less label effort and allows the input of domain knowledge to solve the problems with legal forms such as the "gGmbH".

Hybrid: Rule-based with Machine Learning

The Hybrid approach consists of a rule-based and a supervised ML component to perform the classification and extraction of legal forms from the company name. In the past, rule-based systems were used for the classification of texts. Today, ML approaches are increasingly used. The rules of the rule-based approaches need to be set up manually, which often results in high effort and complexity [32, 33]. In contrast, supervised ML algorithms enable the automated creation of complex sets of rules based on massive amounts of data. However, the algorithms require sufficiently labeled data to learn the rules, which is a one-time manual effort. One advantage of rule-based approaches is that humans can apply their domain knowledge directly when creating a set of rules. This makes the set of rules easy to understand and extensible for humans [32]. The legal form classification and extraction problem demonstrate that legal forms' inconsistent representation and diversity require special domain knowledge. While analysing and labeling the data, we identified other legal forms such as "EK" or "Stiftung", which are not implemented in the rule-based approaches Bundesbank (section 3.1) and Cleanco (section 3.2). Also, we identified other representations for the individual legal forms such as "g.m.b.h." or "o.h.g." that are not implemented in the existing approaches. To solve the classification and extraction problem of legal forms, we combine rule-based components and ML methods to take advantage of both. For this purpose, we divide the legal form classification and extraction problem into the subtasks: (1) identification of legal form relevant tokens, (2) classification of the legal form based on the legal form relevant tokens, and (3) extraction of the legal form relevant tokens

from the company name. The data flow and the solution approach for the hybrid approach's subtasks are shown in figure 3. They are described below:

(1) Identification of legal form relevant tokens: For the legal form's classification, only the legal form tokens in the company name are relevant. For the company name "Example gesellschaft mbh", these are the tokens "gesellschaft" and "mbh". Since we have already established that the diversity of existing legal forms and the inconsistent representation of the individual legal forms requires domain knowledge, we implemented an identification rule set to solve this subtask. The rule set consists of a list of all tokens that are part of a legal form, such as "ek", "eg", "ag" "aktiengesellschaft" or "gmbh". Experts can easily extend this list. With the list's help, all tokens relevant to the legal form of a company name are extracted. The legal form is classified based on the extracted tokens relevant to the legal form in the next step.

(2) Classification of the legal form based on the legal form relevant tokens: For the classification of the legal form based on the extracted legal form relevant tokens we use ML approaches, since the manual creation of a rule set would be very complex and time consuming.

We used and compared the ML methods of Random Forest Tree and Support Vector Classifier (SVC). The labeled data set of the DL approach (section 3.3) is used as training data. The dataset was extended by 500 samples, in which no legal form is included in the company name. Also, the represented legal form was extracted from the BIO labels as a single label. The entire data set thus comprises 18800 training samples. The extracted components are encoded by multi-label binarization. This results in a vector that contains a 1 for each recognized component of the created list and a 0 for all others. With this vector, the two methods were trained with 80% of the data and evaluated with 20% of the data. The results are shown in table 4.

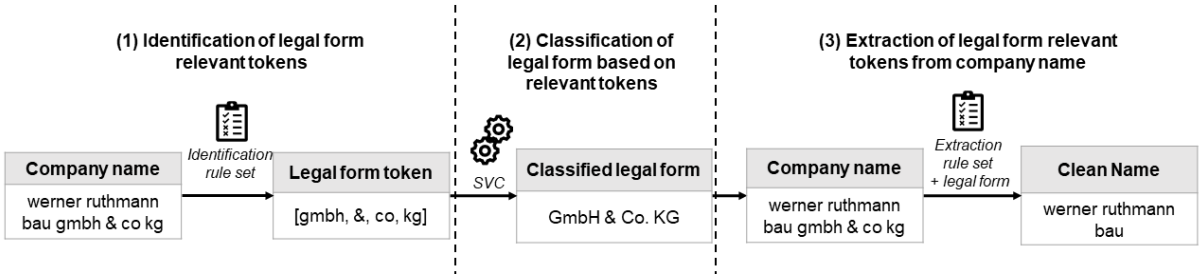

Figure 3: Data flow of the hybrid approach

Table 4 shows that the quality of the models is very high, with over 99%. The SVC shows with a weighted F1 score of 99.7% the better performance than the random forest tree with 99.57%. That both models achieve excellent results shows that the classification of a legal form from the extracted legal form relevant tokens works very reliably. However, the most correct and complete extraction of the legal form components from the company name is decisive for the good performance, which has a significant influence on the classification's success. This shows that combining a rule-based extraction of the legal form relevant tokens and the classification with an ML method is very successful for this application. Finally, we have chosen the SVC as the classifier of the hybrid approach.

Table 4: Results Models for legal form classification

Model	Precision	Recall	F1-Score
Random Forest	0.9963	0.9952	0.9957
SVC	0.9969	0.9971	0.9970

Furthermore, this approach's extensibility is easy to implement by domain experts extending the list of legal-form-relevant tokens. Besides, the effort for labeling new data sets is less than with the DL approach, since it not necessary to label according to the BIO tagging scheme, but rather it is sufficient to label only the legal form belonging to a company name.

(3) Extraction of the legal form relevant tokens from the company name: A rule-based approach does the extraction of the legal form relevant tokens. We have maintained an extraction rule set containing a list of tokens for each legal form that should be searched for and removed from the company name. The rule-based approach needs the previously classified legal form as input. The dependence of the extraction on the classified legal form is an essential condition for the hybrid approach. For example, the following company, "Meyer Gesellschaft Stiftung" is classified as a "Stiftung". In this case, the token "gesellschaft" belongs to the company name and is not a legal form relevant token. If all current legal form relevant tokens would be extracted during the extraction, the token "gesellschaft" beside the "Stiftung" would be erroneously removed from the company name. In our approach, the tokens to be removed depending on the classified legal form. Thus, we ensure that only the tokens belonging to the legal form are removed. In our example, the token “Gesellschaft” is not included in the extraction rule set for the legal form “Stiftung”, so only the token "Stiftung" is removed.

Benchmark of the Approaches

We have created a new labeled data set out of our data sources (see table 1) containing 3733 company names (see table 5). This dataset is used to benchmark the four different approaches for classifying and extracting the company's legal form. This data set is unknown for all approaches to evaluate the quality of the four approaches.

Benchmark data set

When creating a real data set for evaluation, we made sure that the evaluation data set does not contain any company names that have already been used for training the approaches. It was also essential for us to create a real evaluation data set and ensure that all legal forms appear in the data set. Our final dataset contains 3733 company names. These were manually labeled again with the defined BIO tagging scheme (see table 3) in the labeling tool *doccano* [31]. The frequency of each legal form in the evaluation dataset is shown in table 5 (column amount).

Table 5: Benchmark results for every approach with the exact match ratio

		Classification				Extraction		
Legal form	**Amount**	Bundesbank	Cleanco	DL	Hybrid	Cleanco	DL	Hybrid
OHG	345	0.919	0.971	0.980	**0.997**	0.910	**0.962**	0.933
GmbH	324	0.919	0.809	0.969	**0.997**	0.895	**0.966**	0.935
GmbH & Co. KG	307	0.694	0.000	0.958	**0.977**	0.697	**0.945**	0.926
No legal form	287	0.526	**0.969**	0.892	0.937	**0.969**	0.857	0.937
Aktiengesellschaft	257	0.743	0.728	0.969	**0.981**	0.938	0.961	**0.965**
EG	247	0.194	0.854	0.988	**1.000**	0.883	**0.980**	0.960
SE	237	0.958	**1.000**	**1.000**	0.992	0.987	**0.996**	0.987
EK	220	0.000	0.768	0.959	**0.991**	0.900	**0.995**	0.973
Stiftung	216	0.000	0.977	0.958	**0.972**	0.403	0.958	**0.972**
EV	202	0.787	0.832	0.842	**0.990**	0.787	**0.911**	0.896
GbR	198	0.727	0.899	0.934	**1.000**	0.854	**0.949**	0.939
UG	153	0.850	0.935	0.980	**0.987**	0.046	**0.980**	0.974
VVaG	133	0.023	0.571	0.609	**0.774**	0.609	0.602	**0.632**
UG & Co. KG	103	0.660	0.000	0.971	**0.990**	0.000	**0.893**	0.796
KG	84	0.905	0.929	0.976	**1.000**	0.929	0.929	**0.976**
GmbH & Co. KGaA	81	0.790	0.000	0.852	**0.877**	0.000	**0.889**	0.864
gGmbH	79	0.000	0.443	0.418	**0.835**	0.468	0.418	**0.810**

PartG	73	0.000	0.178	0.589	**0.849**	0.096	0.452	**0.521**
GmbH & Co. OHG	59	0.814	0.000	0.831	**0.932**	0.000	0.797	**0.831**
SE & Co. KG	49	0.000	0.000	**1.000**	0.959	0.000	**0.980**	0.959
Stiftung & Co. KG	28	0.000	0.000	**1.000**	**1.000**	0.036	**1.000**	0.929
KGaA	22	0.818	**0.864**	**0.864**	**0.864**	**0.818**	**0.818**	**0.818**
AG & Co. KGaA	11	**0.727**	0.000	**0.727**	**0.727**	0.000	**0.818**	**0.818**
Limited & Co. KG	6	0.833	0.000	0.833	**1.000**	0.000	**0.833**	**0.833**
SE Co. KGaA	5	0.000	0.000	**0.800**	**0.800**	0.000	**0.800**	**0.800**
AG & Co. KG	4	0.750	0.000	**1.000**	0.750	0.000	**1.000**	0.750
AG & Co. OHG	2	**1.000**	0.000	**1.000**	**1.000**	0.000	**1.000**	**1.000**
SE & Co. OHG	1	**1.000**	0.000	0.000	0.000	0.000	0.000	0.000
Summary	3733	0.589	0.696	0.919	**0.962**	0.705	0.913	**0.916**

Execution and analysis of the benchmark

The result of the benchmark is shown in table 5. It was divided into legal form classification (classification of the correct legal form) and extraction (were all legal form components identified in the company name). The Bundesbank approach (section 3.1) is only included in evaluating the classification, as it does not extract the legal form. The benchmark was performed for each legal form. The ratio of correctly classified companies to the total number of companies per legal form was calculated as the so-called exact match ratio.

The exact match ratio for the classification was calculated using the ratio of correctly classified companies to the total number of companies per legal form. The legal form's extraction was evaluated as correct if all legal form elements were extracted from the company name. For the extraction, the exact match ratio was thus calculated from the ratio of correctly extracted legal forms to the respective total number of companies.

Overall, the hybrid approach for the classification and extraction of the legal form is the best of the four approaches. It achieves an exact match ratio of 0.962 for classification and 0.916 for extraction. The DL approach is slightly worse. With an exact match ratio for the classification of 0.919, the DL approach is 4.3% behind the Hybrid approach. For extraction, the DL approach is only 0.3% behind the Hybrid approach. The Bundesbank approach achieved an exact match ratio of 0.589 for the classification. Cleanco achieved an exact match ratio of 0.696 for the classification. For the extraction, Cleanco is with an exact match ratio of 0.705 over 20% behind the DL and Hybrid approach.

In general, the Hybrid approach has a 2-3% better Exact Match Ratio per legal form than the DL approach for the task classification. For the legal forms "EV" or "gGmbH" the Hybrid approach has a 14.8% and 41.7% better exact match ratio. The DL approach often classifies the legal form "EV" as "No legal form" or "EK", which results in a difference of 14.8% of the Exact Match Ratio. For the "gGmbH" legal form, the DL approach often classifies a "GmbH", which results in the 41.7% worse exact match ratio. With the legal forms "SE" and "SE & Co. KG", the DL approach has a 0.8% and 4.1% better exact match ratio than the hybrid approach. In some cases, the hybrid approach classifies an "SE" as "SE & Co. KG" and vice versa, which results in the difference of the exact match ratio. The Bundesbank and Cleanco approach only achieve for the legal forms "KGaA", "AG & Co. KGaA" and "AG & Co. OHG" the same exact match ratio as the DL or hybrid approach. The Bundesbank and Cleanco approaches' rules do not reflect the diverse representations within a legal form to the same extent as the DL and Hybrid approaches. From this, it can be concluded that the legal forms "KGaA", "AG & Co. KGaA" and "AG & Co. OHG" do not show a high diversity in the evaluation data since the approaches have the same exact match ratio. The legal form "Stiftung" is represented very consistently in the evaluation data, as the Cleanco approach has the best exact match ratio of 0.977. The hybrid approach has a 0.5% lower exact match ratio for the same legal form. For the label "No legal form" Cleanco has a 3.2% better exact match ratio than the Hybrid approach. The Cleanco approach covers fewer legal form

variants (see table 3). Cleanco generally classifies more records as "No legal form", which explains the difference.

In the extraction, the difference in the exact match ratio between the DL approach and the Hybrid approach is 0.3%. The differences in the exact match ratio per legal form are minimal as well. In 14 cases, the exact match ratio of the DL approach is better than the Hybrid approach. In 7 cases, the Hybrid approach is better than the DL approach. In 5 cases, the exact match ratios of the two approaches are equal. For the legal form "gGmbH", the exact match ratio difference between the DL approach and the Hybrid approach is 39.2%, which is significantly higher than the others. The DL approach classifies some records with the legal form "gGmbH" as "GmbH" and therefore does not extract the legal form correctly. As a result, the DL approach has a 39.2% worse exact match ratio. The only case where neither the DL approach nor the Hybrid approach has the best exact match ratio is for the label "No legal form". For this label, Cleanco shows the best exact match ratio with 0.969.

Discussion

Theoretical and Practical Implications

The results of our benchmark, which approach is suitable for classifying the company legal form and extracting it from the company name, directly influences theory and practice. First, our developed Hybrid approach increases the data quality of company names and company legal forms in company databases. Our application example for our developed data preparation approach is the company entity matching. Here we show with our research that the classification and extraction of the company legal form is a general problem and exists in many data sources. So far, no benchmark dataset for company entity matching exists which contains this RL challenges. We show that this problem can be solved with our Hybrid approach consisting of a set of rules and a supervised ML method, or our DL approach. Thus, we confirm and support the statements of Govind *et al.* [8] and Gschwind *et al.* [16] that ML procedures should be used for subtasks in RL and thus support the automation of the RL process. This statement is confirmed by our approach and encourages us to identify further general problems in RL and data preparation and investigate suitable ML solutions for these problems. For example, the standardization and matching of company address data. Furthermore, the results of our paper show that it is appropriate for RL to consider problems for the respective real-world entities such as products, persons or companies.

Our developed data preparation approaches, Hybrid and DL, can be used for any new scientific and company data source. We show that general data quality problems with the concise and consistent representation of attributes could be solved with such approaches. The approach could be further explored theoretically and practically and applied to other attributes with similar data quality problems as concise and consistent representation.

Limitations

Our research has limitations that lead to potential future research opportunities. In our paper, we focus on German legal forms. Further research should investigate the extension of the DL and Hybrid approach to include other legal forms. In doing so, the extendibility requirement and performance should be measured. A different approach may be necessary for each country and its legal forms in the future.

We have selected the listed data sources due to our focus on German legal forms (table 1). Future research should investigate additional data sources and additional evaluation data sets. The evaluation of the hybrid and DL approach with other data sources could provide further insights into which approach is the better one under which conditions.

In the future, the approach should be also used in real RL experiments to investigate how much influence this data preparation procedure has on company entity matching results. In addition, a benchmark dataset for company entity matching should be created in order to benchmark existing RL approaches.

Conclusion and Outlook

The entity company is present in many internal and external data sources and is often required in analytical use cases. Therefore, the different internal and external data sources need to be integrated. The integration of the data sources is enabled by the data integration process, which consists of the process steps (1) schema matching, (2) record linkage (RL), and (3) data fusion [7]. In this paper, we focus on the RL of the real-world entity company and define this as company entity matching. In company entity matching, the company name is crucial and presents several challenges. The legal form is often included in the company name and is also an important discriminative attribute. Since the legal form is not a separate attribute in most data sources, it cannot be directly analysed for further data processing steps.

Moreover, the legal form lacks data quality, as it is often not concise and consistent represented in the company name. For the German legal form "GmbH" we show 9 different representations (see table 2). Our goal to solve the data quality problems is to classify and harmonize the legal form and to split the company name and legal form into two separate attributes. To achieve this goal, we answer the following research question in our paper*: "Which approaches are suitable to automatically classify and extract the legal form in the company name?"*. We answer the research question through our inductive data-driven research procedure, according to Grover and Lyytinen [14]. As a result, we have iteratively developed four approaches to solve the problem, which we present and evaluate in a summarising benchmark. The first approach, called Bundesbank, is rule-based and is adapted by the paper by Schild and Schultz [15]. The second approach, called Cleanco, is also rule-based and is adapted on the Github project Cleanco. The third approach, called Deep Learning (DL), defines the legal form classification and extraction problem as a sequence labeling problem and solves it with a Bi-LSTM deep learning model. The fourth approach, called Hybrid, is a combination of a rule set for identification and extraction of legal form relevant tokens and a supervised ML algorithm for the classification of the legal form. The benchmark data set contains 3733 records. The Hybrid approach achieves the best values in the benchmark with an exact match ratio of 96.2% for the legal form classification and 91.6% for the legal form extraction. The DL approach achieved the second-best values with 91.9% for classification and 91.3% for extraction. Thus, the Hybrid shows the best performance in the benchmark. Further, experts can easily extend the developed rule sets, meaning the Hybrid approach is easier to expand than the DL approach. Likewise, additional training data sets can be labeled with new legal forms to extend the classification model. The labeling of new training data sets for the DL approach is more complex since all tokens of the company name must be labeled. In contrast, the supervised ML method in the Hybrid approach requires only one label for the company name.

Our approach and results show that general problems exist for the individual real-world entities such as companies represented in different data sources. For these general-entity-specific problems, generic solutions can be created to improve the data quality, such as concise and consistent representation of attributes. Furthermore, our results show that using hybrid ML methods or DL approaches is successful for these problems and should be further researched. In future research, the developed data preparation approach will be used in RL processes to measure the impact in company based RL case studies.

References

[1] A. Abbasi, S. Sarker, and R. Chiang, “Big Data Research in Information Systems: Toward an Inclusive Research Agenda,” *JAIS*, vol. 17, no. 2, pp. I–XXXII, 2016, doi: 10.17705/1jais.00423.

[2] C. Heinrich and G. Stühler, “Die Digitale Wertschöpfungskette: Künstliche Intelligenz im Einkauf und Supply Chain Management,” in *Fallstudien zur Digitalen Transformation :*

Case Studies für die Lehre und praktische Anwendung, Wiesbaden, Germany: Springer Gabler, 2018, pp. 77–88. https://doi.org/10.1007/978-3-658-18745-3_4

[3] M. Stonebraker and I. Ilyas, "Data Integration: The Current Status and the Way Forward," *IEEE Data Eng. Bull.*, vol. 41, no. 2, 3-9, 2018.

[4] P. Christen, "Data Linkage: The Big Picture," *Harvard Data Science Review*, 2019, doi: 10.1162/99608f92.84deb5c4.

[5] F. Kruse, C. Schröer, and J. Marx Gómez, "Data Source Selection Support in the Big Data Integration Process - Towards a Taxonomy," in *Internationale Tagung Wirtschaftsinformatik (WI)*, Universität Duisburg-Essen, 2021.

[6] X. L. Dong and D. Srivastava, "Big Data Integration," *Synthesis Lectures on Data Management*, vol. 7, no. 1, pp. 1–198, 2015, doi: 10.2200/S00578ED1V01Y201404DTM040.

[7] P. Christen, *Data Matching*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, 10.1007/978-3-642-31164-2

[8] Y. Govind *et al.,* "Entity Matching Meets Data Science: A Progress Report from the Magellan Project," 2019, https://doi.org/10.1145/3299869.3314042

[9] N. Barlaug and J. Atle Gulla, "Neural Networks for Entity Matching: A Survey," 2020, arXiv:2010.11075

[10] Y. Govind *et al.,* "Cloudmatcher: a hands-off cloud/crowd service for entity matching," *Proc. VLDB Endow.*, vol. 11, no. 12, pp. 2042–2045, 2018, doi: 10.14778/3229863.3236255.

[11] P. Christen and W. E. Winkler, "Record Linkage," in *Encyclopedia of Machine Learning and Data Mining*, C. Sammut and G. I. Webb, Eds., Boston, MA: Springer US, 2016, pp. 1–10.

[12] H. Köpcke, A. Thor, S. Thomas, and E. Rahm, "Tailoring entity resolution for matching product offers," in *Proceedings of the 15th International Conference on Extending Database Technology - EDBT '12*, Berlin, Germany, 2012, p. 545.

[13] P. Behnen, F. Kruse, and J. Marx Gómez, "Enhancement of Record Linkage by Using Attributes containing Natural Language Text," in *AAAI-MAKE 2021 Combining Machine Learning and Knowledge Engineering*, Stanford University, Palo Alto, California, USA, 2021, pp. 1–14.

[14] V. Grover and K. Lyytinen, "New State of Play in Information Systems Research: The Push to the Edges," *MISQ*, vol. 39, no. 2, pp. 271–296, 2015, doi: 10.25300/MISQ/2015/39.2.01.

[15] C.-J. Schild and S. Schultz, "Linking Deutsche Bundesbank Company Data using Machine-Learning-Based Classification," 2017, doi: 10.1145/2951894.2951896.

[16] T. Gschwind, C. Miksovic, J. Minder, K. Mirylenka, and P. Scotton, "Fast Record Linkage for Company Entities," in *2019 IEEE International Conference on Big Data (Big Data)*, Los Angeles, CA, USA, 2019, pp. 623–630., 10.1109/BigData47090.2019.9006095

[17] R. Y. Wang and D. M. Strong, "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996, doi: 10.1080/07421222.1996.11518099.

[18] N. Gali, R. Mariescu-Istodor, D. Hostettler, and P. Fränti, "Framework for syntactic string similarity measures," *Expert Systems with Applications*, vol. 129, pp. 169–185, 2019, doi: 10.1016/j.eswa.2019.03.048.

[19] W. Maass, J. Parsons, S. Purao, V. C. Storey, and C. Woo, “Data-Driven Meets Theory-Driven Research in the Era of Big Data: Opportunities and Challenges for Information Systems Research,” *JAIS*, pp. 1253–1273, 2018, doi: 10.17705/1jais.00526.

[20] F. Kruse, A. P. Hassan, J.-P. Awick, and J. Marx Gómez, “A Qualitative Literature Review on Linkage Techniques for Data Integration,” in *53nd Hawaii International Conference on System Sciences, HICSS 2020, Grand Wailea, Maui, Hawaii, USA, January 7-10, 2020*, 2020, pp. 1063–1073. [Online]. 10.24251/HICSS.2020.132

[21] Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan, “Deep Entity Matching with Pre-Trained Language Models,” arXiv:2004.00584, 2020.

[22] S. Mudgal *et al.,* “Deep Learning for Entity Matching,” in *Proceedings of the 2018 International Conference on Management of Data - SIGMOD '18*, Houston, TX, USA, 2018, pp. 19–34.

[23] M. Ebraheem, S. Thirumuruganathan, S. Joty, M. Ouzzani, and N. Tang, “Distributed representations of tuples for entity resolution,” *Proc. VLDB Endow.*, vol. 11, no. 11, pp. 1454–1467, 2018, doi: 10.14778/3236187.3236198.

[24] A. Doan *et al.,* “Magellan: Toward Building Ecosystems of Entity Matching Solutions,” *Commun. ACM*, vol. 63, no. 8, pp. 83–91, 2020, doi: 10.1145/3405476.

[25] J. Cuffe and N. Goldschlag, “Squeezing More Out of Your Data: Business Record Linkage with Python,” in 2018.

[26] S. M. Randall, A. M. Ferrante, J. H. Boyd, and J. B. Semmens, “The effect of data cleaning on record linkage quality,” *BMC medical informatics and decision making*, vol. 13, pp. 1–10, 2013, doi: 10.1186/1472-6947-13-64.

[27] I. Koumarelas, L. Jiang, and F. Naumann, “Data Preparation for Duplicate Detection,” *Journal of Data and Information Quality (JDIQ)*, vol. 1, no. 1, pp. 1–24, 2020, https://doi.org/10.1145/3377878

[28] A. Akhundov, D. Trautmann, and G. Groh, “Sequence Labeling: A Practical Approach,” *CoRR*, arXiv:1808.03926, 2018.

[29] S. Liu, B. Tang, Q. Chen, and X. Wang, “Drug Name Recognition: Approaches and Resources,” *Information*, vol. 6, no. 4, pp. 790–810, 2015, doi: 10.3390/info6040790.

[30] X. Zhong, E. Cambria, and A. Hussain, “Extracting Time Expressions and Named Entities with Constituent-Based Tagging Schemes,” *Cogn Comput*, vol. 12, no. 4, pp. 844–862, 2020, doi: 10.1007/s12559-020-09714-8.

[31] H. Nakayama, T. Kubo, J. Kamura, Y. Taniguchi, and X. Liang, *doccano: Text Annotation Tool for Human.* [Online]. Available: https://github.com/doccano/doccano

[32] Julio Villena Roman, Sonia Collada-Perez, Sara Lana-Serrano, and Jose C. González-Cristobal, “Hybrid Approach Combining Machine Learning and a Rule-Based Expert System for Text Categorization,” 2011.

[33] Fabrizio Sebastiani, “Machine Learning in Automated Text Categorization,” *ACM computing surveys (CSUR)*, 2002. [Online], https://doi.org/10.1145/505282.505283 1