Case-based Data Masking for Software Test Management

Mirjam Minor[1], Alexander Herborn[2], and Dierk Jordan[2]

[1] Goethe University, Business Information Systems, Robert-Mayer-Str. 10, 60629 Frankfurt, Germany
[2] R + V Versicherung AG, Wiesbaden, Germany
minor@cs.uni-frankfurt.de, Alexander.Herborn@ruv.de, Dierk.Jordan@ruv.de

Abstract. Data masking is a means to protect data from unauthorized access by third parties. In this paper, we propose a case-based assistance system for data masking that reuses experience on substituting (pseudonymising) the values of database fields. The data masking experts use rules that maintain task-oriented properties of the data values, such as the environmental hazards risk class of residential areas when masking address data of insurance customers. The rules transform operational data into hardly traceable, masked data sets, which are to be applied, for instance, during software test management in the insurance sector. We will introduce a case representation for masking a database column, including problem descriptors about structural properties and value properties of the column as well as the data masking rule as the solution part of the case. We will describe the similarity functions and the implementation of the approach by means of myCBR. Finally, we report about an experimental evaluation with a case base of more than 600 cases and 31 queries that compares the results of a case-based retrieval with the solutions recommended by a data masking expert.

Keywords: CBR applications, data protection, substitution rules, myCBR

1 Introduction

A novel General Data Protection Regulation (GDPR) [1] is effective in the European Union from May 25, 2018. Organizations in non-compliance may face heavy fines. The regulation document states that processing of personal data is lawful if the data subject has given consent [1, p.36]. It recommends encryption or pseudonymisation in cases where the processing is for a purpose that is different from that for which the personal data have been collected [1, p.37]. Apart from the fact that there remains a grey zone which purposes are still based on the data subject's consent, it is quite challenging for organizations to follow this recommendation.

Encryption is not feasible for many purposes of data processing, such as data analytics, business process modeling, or software testing, that require unencrypted data with values as realistic as possible. *Pseudonymisation* means "the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information" [1, p.33]. In contrast to pseudonymisation, anonymisation produces data where the personal data is not traceable at all any more [4]. Nulling-out is a sample anonymisation technique. A straight-forward technique of pseudonymisation is to replace data values by arbitrary pseudonyms, for instance, transforming the name Bob into Alice. However, this substitution may lead to a severe loss of information, including task-oriented properties of the data that are necessary to fulfill the intended purpose of data processing. An example is the approximate age of a policyholder that is important for the purpose of testing software in an insurance company. In such scenarios, the date of birth should not be pseudonymised arbitrarily. More sophisticated techniques of pseudonymisation are called data masking. *Data masking* is "a technique applied to systematically substitute, suppress, or scramble data that call out an individual, such as names, IDs, account numbers, SSNs (i.e. social security numbers, editor's note), etc." [7, p.8]. Data masking rules can be specified by data masking experts to identify and transform personal data into pseudonymised target data. This is a time-consuming task that has to be conducted carefully each time a new data source is pseudonymised.

Case-based reasoning (CBR) provides a means to reuse experience in data masking. In this paper, we present a case-based assistance system for specifying masking rules. We demonstrate the feasibility of the approach in the application area of test management for insurance software.

The remainder of the paper is organized as follows. The basic principles and techniques of data masking are introduced in Section 2. The case representation is presented in Section 3. Section 4 sketches the retrieval with the according similarity functions. In Section 5, the approach is evaluated in a lab experiment. Section 6 contains a discussion and a conclusion.

2 Data masking

A data masking rule transforms personal data into hardly traceable, masked data. In software test management, the source for data masking are operational data with newly created test data sets as target. Following Raghunathan [4, p.172ff], there are various basic techniques for data masking that reduce the information content of the data to different degrees. Deterministic techniques achieve reproducible results. Randomised techniques produce different results for each run. The personal data used by the latter is traceable only by means of log information on the masking process, which blurs the line between pseudonymisation and anonymisation.

Substitution rules replace a data value by another data value. The following variants of substitution are particularly useful for pseudonymisation purposes:

- *Direct substitution* maps a data value directly to a substitute value.
- *Lookup substitution* uses an external list of potential replacements. In lookup substitution with hashing, the list is organised in a hash table. A mapping function assigns a hash-value to each source data value. In contrast, randomised lookup substitution selects a value from the list arbitrarily.
- *Conditional substitution* is an extension of direct or lookup substitution that considers conditions.

Obviously, direct substitution and lookup substitution with hashing are deterministic while randomised lookup substitution is nondeterministic. An example for a conditional substitution is a lookup substitution that maintains a risk class of the data values, such as the environmental hazards risk class of residential areas when masking address data of customers.

Shuffling rules rearrange the data within the same column across different rows. Several data columns can be grouped in order to preserve the relationship of their values. Table 1 depicts an example where postal code, city and street are grouped together during shuffling.

Table 1. Example for data masking by shuffling.

Raw data

	Customer id	Postal code	City	Street
1	10012	65189	Wiesbaden	Siegfriedring
2	10049	65195	Wiesbaden	Lahnstr.
3	10144	55122	Mainz	Saarstr.
4	10220	60486	Frankfurt	Solmsstr.
5	13002	60594	Frankfurt	Dreieichstr.

Masked data

	Customer id	Postal code	City	Street
1	10012	60594	Frankfurt	Dreieichstr.
2	10049	55122	Mainz	Saarstr.
3	10144	60486	Frankfurt	Solmsstr.
4	10220	65189	Wiesbaden	Siegfriedring
5	13002	65195	Wiesbaden	Lahnstr.

Mutation rules produce variations of data values within given boundaries. In case of numeric values, the number variance technique exchanges a number by a randomly generated value within a range. An advanced variance technique randomises the selection of the arithmetic operator or function to mutate the numeric data value. For instance, the mutation of a date value can process the date for a random number x between 0 and 10 as follows:

- $0 \leq x \leq 3$: increase date by 30 days
- $4 \leq x \leq 7$: decrease date by 70 days
- $8 \leq x \leq 10$: increase date by 120 days

A birth date 06/01/1955 is mutated to 03/23/1955 in case x is 5.

The data masking rules are used by an ETL[3] tool within a data masking architecture (see Figure 1). The source data comprises of database tables. During

[3] ETL stands for Extract - Transform - Load

export, the data are split into tables with the portions of data that contain personal data and tables without personal data. The tables with personal data are loaded into staging data for the ETL tool. The ETL tool applies the masking rules and exports the masked data to be loaded into the target data.

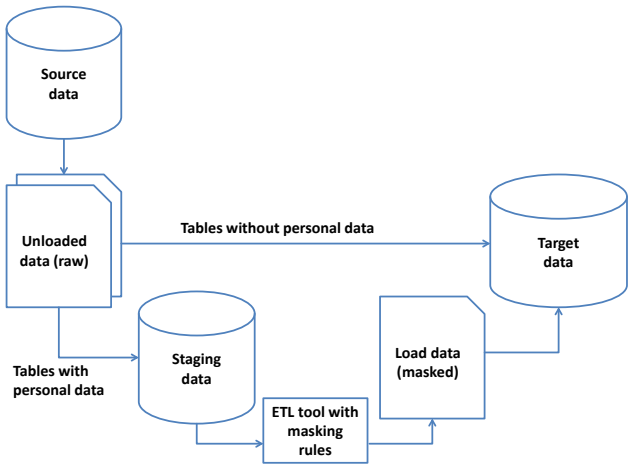

Fig. 1. Architecture for data masking.

3 Representation of a data masking case

A data masking case represents the experience on masking one attribute of a database, i.e. it is used to replace all values of a particular data column in a database table. The problem description comprises a set of descriptors for the data column. The solution is a masking rule that is used to pseudonymise the values of the data row. In addition to the attribute to be masked, the masking rule might require access to several attributes of the database table.

A sample rule for masking the first names while maintaining the gender might use lookup substitution with hashing as follows. The rule expects two attributes namely the salutation and the first name as input. It creates a hash value from the input serving as key for a gender-specific lookup table with first names. It selects the element with the key as pseudonym. Table 2 depicts six sample cases for different attributes. Each row stands for a case with the problem part on the left hand side and the solution part (the masking rule) on the right hand side. Case 2 employs the sample rule described above as the solution for masking the first names.

The problem descriptors of a case represent structure-oriented and value-oriented properties of the data column. In the sample case base (see Table 2),

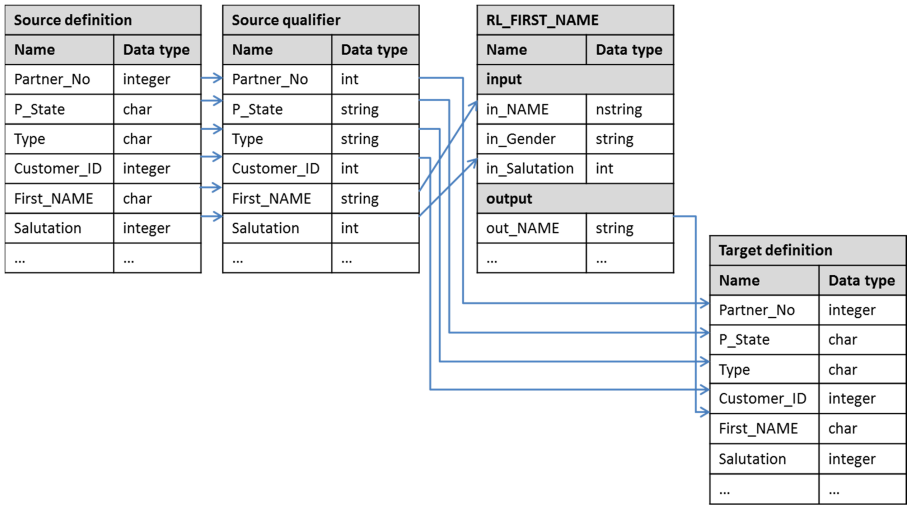

Fig. 2. Sample mapping for the data masking rule RL_FIRST_NAME.

the structure-oriented properties are the column name and the data type of the attribute. The value-oriented properties are the most used value in the column, the frequency of the most used value in percentage of occurrence (value frequency), a regular expression describing the most used pattern, the frequency of this pattern (pattern frequency), and a value that describes the percentage of data values with unique occurrence within the data row (distinct percentage). Michael, for instance, is the most used value in cases 2 and 6. 'XXXXXX' stands for a sequence of six letters. Since Michael has 7 letters, it does not match the most used pattern in cases 2 and 5.

Table 2. Sample cases for different data rows.

	Column name	Data type	Most used value	Value frequency (in %)	Most used pattern	Pattern frequency (in %)	Distinct percentage (in %)	Rule
1	Name	string	Müller	0.54	XXXXXX	17.88	13.86	RL_NAME
2	First_Name	string	Michael	1.2	XXXXXX	20.81	2.05	RL_FIRST_NAME
3	Street	string	Hauptstr.	1.8	XXXXXXXX.	5.68	2.52	RL_ADDRESS_V6
4	TaxID	string	n.d.	0.08	999999999	7.23	48.93	RL_TAXID_V_STR
5	BIC	int	0	10.76	99999999	89.24	21.48	RL_BIC_V5
6	Prename	string	Michael	0.93	XXXXXX	15.76	2.46	RL_FIRST_NAME

4 Retrieval of data masking cases

The case-based system provides assistance to data masking experts in specifying data masking rules. The data table to be pseudonymised is processed automatically resulting in n queries where n is the number of columns. The query descriptors can be easily determined from the data source by means of SQL queries.

The case-based system performs a retrieval process providing the best matching case for each query. The system uses standard similarity measures [2] as local similarity functions (compare Table 4). The percentage values are compared by a linear function. 'Most used value', 'Most used pattern' and 'Column name' are compared by string matching. 'Data type' uses a symbolic similarity function depicted in Table 3.

Table 3. Sample similarity values for data types.

	date/time	decimal	int	string
date/time	1.0	0.0	0.0	0.0
decimal	0.0	1.0	0.5	0.0
int	0.0	0.5	1.0	0.0
string	0.0	0.0	0.0	1.0

The values of the local similarity functions are aggregated by a weighted sum [5, p.29]:

$$\sum_{i=1}^{i=n}\left(\frac{\omega_i sim(x_i,y_i)}{k}|1 \leq i \leq n\right).$$

The weights ω_i for the local similarity values have been specified as depicted in Table 4. The weights are preliminary and intend to estimate the impact of a descriptor for the result. In the recent specification, the value frequency and the pattern frequency are not considered during retrieval.

Table 4. Types and weights of the local similarity functions.

Column name	Data type	Most used value	Value frequency	Most used pattern	Pattern frequency	Distinct percentage
String matching	Symbolic	String matching	Linear	Partial string matching	Linear	Linear
$\omega_1 = 5$	$\omega_2 = 3$	$\omega_3 = 2$	$\omega_4 = 0$	$\omega_5 = 4$	$\omega_6 = 0$	$\omega_7 = 1$

5 Evaluation

The approach is implemented by means of the myCBR tool [6]. The evaluation provides a first proof-of-concept that is guided by the following two hypotheses:

H1 The CBR approach is able to identify columns with personal data in database tables.

H2 The retrieval result provides a solution that is comparable to the data masking rule that is recommended for a column by a human expert.

The evaluation uses an experimental setting with real sample data from the insurance domain. 615 cases have been extracted from an operational database system of the insurance company R + V with the according data masking rules specified by experts from the test management department.

Overall, R + V uses a heterogeneous landscape of database systems, including IBM DB2, Microsoft SQL Server, SAP R/3 and Hana. Today, the experts sift all tables when preparing test data. The tables that include personal data are selected as staging data (cmp. Figure 1) for a detailed review. During review, the experts mark columns of the sensitive tables that need not to be masked and specify rules for the other, sensitive columns. The entire staging data created from R + V's IBM DB2 database serves as the source for our case base in the experimental setup. It includes nearly 200 tables that contain personal data for different insurance products, such as health, life, casualty etc. The experimental case base is created automatically from the 615 columns of the staging data tables. The values of the problem descriptors 'Column name', 'Data type', 'Most used value', 'Value frequency', 'Most used pattern', 'Pattern frequency', and 'Distinct percentage' (cmp. Table 2) are extracted by means of the ETL tool Informatica Analyst and by SQL statements. The solution parts of the cases (the masking rules) are taken directly from the ETL tool as they have been specified for each column by the experts. The experimental case base contains 90 'negative' cases on the 90 columns that have been marked as 'not sensitive' by the experts. For instance, the 'tariff rate' of a particular database table has been excluded from masking. 38 different masking rules have been specified for the columns resulting in 525 'positive' cases. In addition to the case base, the experimental setup comprises a set of queries. 31 queries have been created from two database tables that are actually to be pseudonymised for software testing. Further, the similarity functions are specified in myCBR as described in Section 4. In order to investigate the hypotheses H1 and H2, the results of the case-based retrieval are compared to a golden standard provided by the data masking experts for the recent two database tables.

The preparation of the experimental setup from real data allows already some observations with respect to the feasibility of the chosen case representation. As expected, the value-oriented descriptors are useful to understand the experts' rule assignment. Unsurprisingly, some columns with different names refer to matching content and use the same masking rule, such as 'VT_GEB_DAT' and 'GEBURTS_DATUM' for birthdays. In some cases, underspecified data types require slightly different masking rules for matching content, for instance 'string' used in the format 'mm/dd/yy' in one column vs. 'dd.mm.yyyy' in another column. In such cases, the value-oriented descriptors differ. Further, some structure-

oriented descriptors, such as 'SCHL2' as column name, are not comprehensible at all while the value-oriented descriptors provide further explanation and, thus, are justified.

The evaluation results are depicted in Figure 3 including the retrieval results for the 31 queries. Queries T1 to T13 are created from the first database table. T14 to T31 origin from the second database table. The right hand side of the table shows the recommendations of the experts for the queries which serve as the golden standard. For 11 queries, the experts recommend a masking rule. For 9 queries, the values of the similarity function are above a threshold of 0.8. This can be interpreted as a hint that the considered data row contains personal data that is to be masked. The comparison of the result of the prototype with the recommendation by the experts shows that it includes three false negatives for T5, T7, and T9. This is not yet satisfactory. Further, the result contains one false positive (T30), which is not so bad. Thus, hypothesis H1 is partially fulfilled.

Including the false negatives, however, the 100% overlap of the recommendations of the CBR system with the golden standard is surprisingly high. The experiments confirm hypothesis H2.

6 Discussion and conclusion

In this paper, we have introduced data masking as a novel application task for CBR. The case-based assistance system aims at assigning data masking rules to database tables with the goal to mask personal data by pseudonyms. We presented a structural case representation whose problem descriptors are extracted from database columns automatically. We implemented the approach with the tool myCBR for the purpose of software test management in the insurance sector. The results of the preliminary evaluation show that CBR is a promising approach.

The results create manifold opportunities for future work. First, we are planning to conduct further experiments. The weights that specify the relevance of the local similarity functions will be further investigated, for instance from an information theoretical approach. In addition, the dependencies between columns that are already considered in the solution parts (rules) could also serve as contextual information during retrieval. However, masking rules can get quite complex depending on the properties that need to be preserved. For example, the rule RL_ADDRESS_V6 takes customer address data like street name, zip code or city as input and treats it according to the following procedure:

- Check whether the given zip code is valid or not. If not, take the given city name and get the corresponding zip code.
- Use the zip code as a parameter for a lookup to get the information how many streets and ZUERS-entries [4] exist in the lookup-table for that zip code.

[4] ZUERS is a zoning-system that is determined by the potential risk to become victim of a flooding or a similar environmental hazard. The ZUERS-zone is an important criteria to calculate the insurance rate, e.g. of a residence insurance.

- The street name and house number are converted by a deterministic, reproducible hash-algorithm and projected to a number between one and the amount of entries in the lookup-table that have the same zip code and ZUERS-zone as the original value.
- Take the entry of the lookup-table where the index is equivalent to the calculated number and use it to replace the original value.

This procedure ensures that the masked value has the same city and ZUERS-zone as the original value. Thus, a residence insurance at the masked address is guaranteed to cost the same as at the original address. This is required for testing the residence insurance calculation. The rule deals with dependencies between database columns and from external knowledge sources, such as those between zip code, city and ZUERS-code. It requires pre-processing to transform the representation of such dependencies from the rule into contextual information that is accessible during retrieval. Further, the threshold for the global similarity function could be improved or replaced by a machine learning approach that identifies which columns should be masked at all previously to the CBR approach that selects an appropriate masking rule. Larger experiments are required to demonstrate the scalability of the approach for both, an increasing number of database columns and rows. We expect that the number of data entries (rows) can be easily increased since the problem descriptors are determined off-line. A larger number of database columns or entire tables will probably require more sophisticated memory models for the case base.

Reducing the number of persons who have access to personal data means reducing the risk of incidents with data leakage. Today, productive data including personal data are still used also in non-productive systems [3]. The manual efforts for pseudonymisation that maintains task-oriented properties for the non-productive systems are tremendous. CBR has demonstrated a high potential to provide automated assistance for the data masking experts. Decreasing the efforts for pseudonymisation will significantly contribute to the companies' capability to achieve compliance with data protection regulations such as GDPR.

7 Acknowledgements

The authors would like to thank the data masking experts of R + V who contributed to this work by their rule recommendations. Providing the golden standard for the evaluation, they are vitally important to demonstrate the feasibility of the approach. We highly appreciate their time and efforts.

References

1. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Official Journal of the European Union*, L 119, 2016.
2. R. Bergmann. *Experience management: Foundations, development methodology, and Internet-based applications.* Springer Verlag, 2002.

3. A. Lang. Anonymisierung/Pseudonymisierung von Daten für den Test. In *D.A.CH Security Conference 2012*, Konstanz, 2012. Syssec Forschungsgruppe Systemsicherheit.
4. B. Raghunathan. *The Complete Book of Data Anonymization: From Planning to Implementation*. CRC Press, 2013.
5. M. M. Richter and R. Weber. *Case-Based Reasoning: A Textbook*. Springer, auflage: 2013 edition, Nov. 2013.
6. A. Stahl and T. Roth-Berghofer. Rapid Prototyping of CBR Applications with the Open Source Tool myCBR. In K.-D. Althoff, R. Bergmann, M. Minor, and A. Hanft, editors, *Advances in Case-Based Reasoning, 9th European Conference, ECCBR 2008, Trier, Germany, September 1-4, 2008. Proceedings*, volume 5239 of *Lecture Notes in Computer Science*, pages 615–629. Springer, 2008.
7. N. Venkataramanan and A. Shriram. *Data Privacy: Principles and Practice*. CRC Press, 2016.

Test case ID	Column name	Highest similarity value	Most similar column	Rule recommendation by prototype	Rule recommendation by expert
T1	HAUS_NR_ZUSATZ	0.86	Q01T221_MT_HAUS_NR_ZUSATZ	RL_Anschrift_V6	RL_Anschrift_V6
T2	ANREDE	0.79	Q35T806_MT_RUF_NR_ERG	RL_Telefonnummer_V_Text_Robust	none
T3	NATION_KZ	0.87	Q45T503_MT_NATION_KZ	RL_Land_V_Str_Kfz_v2	RL_Land_V_Str_Kfz_v2
T4	HAUS_NR	0.86	Q45T551_MT_UNTERN_HAUS_NR	RL_Anschrift_V6	RL_Anschrift_V6
T5	ORT_ZUS	0.65	Q45T551_MT_UNTERN_ORT	RL_Anschrift_V6	RL_Anschrift_V6
T6	VORNAME	0.85	Q12T120_MT_VORNAME1	RL_Vorname_V2	RL_Vorname_V2
T7	NAME1	0.77	Q55T009_MT_VWB_NAME_1	RL_Nachname_V2	RL_Nachname_V2
T8	STAAT_ZUGEH	0.87	Q01T001_MT_STAAT_ZUGEH	RL_Land_V_Str_Kfz_v2	RL_Land_V_Str_Kfz_v2
T9	STR	0.56	Q80T102_MT_STR	RL_Anschrift_V6	RL_Anschrift_V6
T10	FK_Q01T001UID	0.41	Q45T552_MT_UNTERN_ANPR_TEL01	RL_Telefonnummer_V_Text_Robust	none
T11	GEBURTSDATUM	0.84	Q17T415_MT_GEBURTSDATUM	RL_Geburtstag_V_Dat_V2	RL_Geburtstag_V_Dat_V2
T12	ORT_NAME	0.82	Q17T200_MT_ORT_NAME	RL_Anschrift_V6	RL_Anschrift_V6
T13	PLZ	0.86	Q45T551_MT_UNTERN_PLZ	RL_Anschrift_V6	RL_Anschrift_V6
T14	PR_FAKTOR	0.57	Q12T576_MT_FAX_NR_AKT	RL_Telefon_Num_Vorwahl_Durchwahl	none
T15	BEITR_PRO_A_VTRG	0.54	Q12T576_MT_FAX_NR_AKT	RL_Telefon_Num_Vorwahl_Durchwahl	none
T16	FK_Q35T101RG_TYP	0.63	Q35T811_MT_BTG_FKT_HSNR_ZUS	RL_Anschrift_V6	none
T17	FK_Q35T101RG_TYP_Z	0.65	Q35T811_MT_BTG_FKT_HSNR_ZUS	RL_Anschrift_V6	none
T18	TARIFSATZ	0.55	Q12T576_MT_TEL_VW_TAG	RL_Telefon_Num_Vorwahl_Durchwahl	none
T19	ERF_LFD_NR	0.67	Q35T805_MT_HSNR	RL_Anschrift_V6	none
T20	JAHR_PR	0.56	Q12T516_MT_TEL_VW_VN_PRIV	RL_Telefon_Num_Vorwahl_Durchwahl	none
T21	FK_Q35T101AG	0.62	Q35T811_MT_BTG_FKT_HSNR	RL_Anschrift_V6	none
T22	GRUND_PR	0.59	Q57T305_MT_BERUF_NR	RL_Beruf_in_Unfall_V2	none
T23	H_LFD_NR	0.68	Q35T811_MT_BTG_FKT_HSNR	RL_Anschrift_V6	none
T24	MASCH_UMS_ABR_KZ	0.75	Q17T732_MT_KFZ_AKZ_BUCH	RL_KFZ_V_NUM_Optimiert	none
T25	MIND_PR	0.55	Q12T516_MT_TEL_VW_VN_PRIV	RL_Telefon_Num_Vorwahl_Durchwahl	none
T26	PR_FAKTOR_INTERN	0.53	Q12T576_MT_TEL_NR_AKT	RL_Telefon_Num_Vorwahl_Durchwahl	none
T27	VEREINB_PR_SATZ	0.54	Q35T815_MT_NEG_VER_HSNR	RL_Anschrift_V6	none
T28	FK_Q35T101SACH_VSN	0.54	Q35T122_MT_SICH_BLZ	RL_BIC_IBAN_BLZ_Ktonr_V_BLZ_Ktonr_Num_v5	none
T29	PR_KLASSE	0.7	Q03T110_MT_KORRESPONDENZFELD	ToDo	none
T30	UMS_VOM_DAT_X	0.8	Q17T756_MT_GEB_DAT_VP	RL_Geburtstag_V_Dat_V2	none
T31	UMSATZ_BETR	0.49	Q45T552_MT_UNTERN_KTONR	RL_BIC_IBAN_BLZ_Ktonr_V_BLZ_Ktonr_Num_v5	none

Fig. 3. Evaluation result.