Knowledge-Based Techniques for Document Fraud Detection: A Comprehensive Study

Beatriz Martínez Tornés, Emanuela Boros, Antoine Doucet, Petra Gomez-Krämer, Jean-Marc Ogier, Vincent Poulain D'andecy

Knowledge-based Techniques for Document Fraud Detection: A Comprehensive Study*

Beatriz Martínez Tornés[1][0000−0002−7820−640X], Emanuela Boros[1][0000−0001−6299−9452], Antoine Doucet[1][0000−0001−6160−3356], Petra Gomez-Krämer[1][0000−0002−5515−7828], Jean-Marc Ogier[1][0000−0002−5666−475X], and Vincent Poulain d'Andecy[2]

[1] University of La Rochelle, L3i, F-17000, La Rochelle, France
firstname.lastname@univ-lr.fr
[2] Yooz, 1 Rue Fleming, 17000 La Rochelle, France
firstname.lastname@getyooz.com

Abstract. Due to the availability of cost-effective scanners, printers, and image processing software, document fraud detection is, unfortunately, quite common nowadays. The main challenges of this task are the lack of freely available annotated data and the overflow of mainly computer vision approaches. We consider that relying on the textual content of forged documents could provide a different view on their detection by exploring semantic inconsistencies with the aid of specialized knowledge bases. We, thus, perform an exhaustive study of existing state-of-the-art methods based on knowledge-graph embeddings (KGE) using a synthetically forged, yet realistic, receipt dataset. We also explore additional knowledge base incremental data enrichments, in order to analyze the impact of the richness of the knowledge base on each KGE method. The reported results prove that the performance of the methods varies considerably depending on the type of approach. Also, as expected, the size of the data enrichment is directly proportional to the rise in performance. Finally, we conclude that, while exploring the semantics of documents is promising, document forgery detection still poses a challenge for KGE methods.

Keywords: Fraud Detection · Knowledge Base · Knowledge Graph.

1 Introduction

Document forgery is quite common nowadays due to the availability of cost-effective scanners, printers, and image processing software. Most administrative documents exchanged daily by companies and public administrations lack the technical securing to authenticate them, such as watermarks, digital signatures or other active protection techniques. Document forgery is a gateway to other types of fraud, as it can produce tampered supporting documents (invoices, birth

* This work was supported by the French defense innovation agency (AID) and the VERINDOC project funded by the Nouvelle-Aquitaine Region.

certificates, receipts, payslips, etc.) that can lead to identity or tax fraud. The amount of fraud detected in tax and social matters reached €5.26 billion in 2019, according to the CODAF[3] (French Departmental anti-fraud operational committees). According to Euler Hermes [24], European fraud insurance company, and the French Association of Financial Directors and Management Controllers (DFCG) on their annual fraud trend analysis the most common fraud attempts suffered by companies in France in 2019 are: fake supplier fraud (by 48% of respondents), fake president fraud (38%), other identity theft, e.g., banks, lawyers, auditors etc., (31%) and fake customer fraud (24%).

One of the main challenges of document fraud detection is the lack of freely available tagged data, as many studies around fraud do not consider the actual documents and focus on the transactions (such as credit card fraud, insurance fraud or even financial fraud) [10, 36, 43]. Collecting real forged documents is rather difficult [48, 39, 54], because real fraudsters would not share their work, and companies or administrations are reluctant to reveal their security breaches and cannot share sensitive information. Taking an interest in real documents actually exchanged by companies or administrations is important for the fraud detection methods developed to be usable in real contexts and for the consistency of authentic documents to be ensured.

Most of the recent research in document forensics is focused on the analysis of images of documents as a tampering detection task [13, 19, 25, 17]. Likewise, most of the existing corpora for fraud detection are based on the creation of synthetic content by introducing variations allowing a particular approach to be tested [39, 10, 43]. For example, documents were automatically generated [12], as well as the noise and change in size of some characters, their inclination or their position. A corpus of payslips was artificially created by randomly completing the various fields required for this type of document [48]. Another corpus consists of the same documents scanned by different scanners in order to evaluate source scanner identification [42]. These corpora examples, suitable for fraud detection by image-based approaches [18, 20, 21], are not appropriate for content analysis, because they do not include realistic information, nor frauds that are semantically more inconsistent or implausible than the authentic documents.

However, we consider that the textual content of the document could provide a different vision towards detection fraud that would not rely on graphical imperfections, but on semantic inconsistencies. Existing fraud detection approaches mainly focus on supervised machine learning (e.g., neural networks, bagging ensemble methods, support vector machine, and random forests) based on handcrafted feature engineering [39, 10, 36, 37, 35]. However, these approaches do not consider documents as their only input. Knowledge graph representation methods [56, 32, 50] are used to predict the plausibility of the facts or *fact-checking* using external knowledge bases (YAGO [49], DBpedia [6] or Freebase [14]). Despite the popularity of knowledge graph-based approaches, these are still underexplored in analyzing document coherence and detecting fraud in semi-structured

[3] Comités opérationnels départementaux anti-fraude `https://www.economie.gouv.fr/codaf-comites-operationnels-departementaux-anti-fraude`

administrative documents (receipts, payslips). Semi-structured data is data that presents some regularity, but not as much as relational data [1]. Most studies focused on coherence analysis and approached it from a discursive point of view [9], with tasks such as sentence intrusion detection [46] or text ordering: this approach is not compatible with administrative documents. We, thus, propose to perform an exhaustive study on knowledge-based representation learning and its applicability to document forgery detection, in a realistic dataset regarding receipt forgery [3, 4].

The remainder of the paper is organized as follows. First, Section 2 presents this dataset in detail along with the ontology based on its topical particularity. Section 3 defines the notions of fact-checking and document fraud detection with knowledge bases. Section 4 provides an exhaustive list of the knowledge-based different state-of-the-art methods that further are explored in the experimental setup in Section 5. The results are presented and visualized in Section 6. Finally, results are discussed and conclusions drawn in Section 7.

2 Receipt Dataset for Fraud Detection

The dataset [3, 4] is composed of 999 images of receipts and their associated optical character recognition (OCR) results, of which 6% were synthetically forged. It was collected to provide a parallel corpus (images and texts) and a benchmark to evaluate image- and text-based methods for fraud detection.

Fig. 1. A scan of a normal receipt (left) and a forged receipt (left) from the dataset [4]. The red box reveals the removal of a grocery item.

The forged receipts are the result of forgery workshops, in which participants were given a standard computer with several image editing software to manually alter both images and associated OCR results of the receipts. The workshop fraudsters were free to choose the image modification method. The corpus contains copy-move forgeries (inside the document), splicing (copying from a different document), imitation of font with a textbox tool, etc. Not only are the forgeries diverse, but they are also realistic, consistent with real-world situations such as a fraudulent refund, an undue extension of warranty or false

mission expense reports, as shown in Figure 1. The dataset also provides an ontology [5].The ontology accounts for all the information present in a receipt: the classes that define *Company* entities, through their contact information (telephone numbers, website, address, etc.) or information enabling them to be identified (SIREN[4], SIRET[5]). Other classes concerns purchases (purchased products, by which companies, means of payment, etc.).

3 Knowledge-based Techniques for Document Fraud Detection

Document fraud detection is closely related to *fact-checking* [28]. *Fact-checking* is the NLP task that refers to the evaluation of the veracity of a claim in a given context [51]. We consider that the objectives between fact-checking and document authentication are similar enough to focus on their commonalities and their intersection for the detection of document fraud.

Although tampering, or fraud, can be mentioned as part of a verification or authentication task, this positions tampering as a barrier impacting data quality and not as a characteristic of the data to be detected. Data falsification can be identified as an obstacle to information verification [11] that is resolved by assessing the reliability of sources or by relying on information redundancy (majority voting heuristic).

Recent *fact-checking* methods utilize knowledge bases to asses the veracity of a claim. A *knowledge base* (KB) is a structured representation of facts made up of entities and relations. Entities can represent concrete or abstract objects and relations represent the links maintained between them. The terms *knowledge graph* and *knowledge base* are often used interchangeably [33]. As we are more particularly interested in formal semantics and in the interpretations and inferences that can be extracted from it, we prefer the term *knowledge base*. *Link prediction* consists of exploiting existing facts in a knowledge graph to infer new ones. More exactly, it is the task that aims to complete the triple (subject, relation,?) Or (?, relation, object) [44]. The information extracted from an administrative document can be structured in the form of triples and be considered as a *knowledge base* whose veracity is to be assessed.

4 Knowledge-based Fact-Checking Methods

Many knowledge base fact-checking methods have been proposed, as well as many variations and improvements thereof. First considered as a logic task, the goal was to find and explicitly state the rules and constraints. Those rules were either manually crafted or obtained with rule mining methods [26]. More recently, with the rise of machine learning techniques, knowledge base methods with graph embeddings (KGE) methods have been proposed [55, 33, 31, 44]. These methods

[4] `https://en.wikipedia.org/wiki/SIREN_code`
[5] `https://en.wikipedia.org/wiki/SIRET_code`

encode the entities and the relations between them of the knowledge base in a vector space of low dimensionality. These vectors aim to capture latent features of the entities (and relations) of the graph.

Next, while we provide an exhaustive list of the different methods, we focus on detailing the KGE-based approaches that we further consider in the experimental setup in Section 5. We separated the methods according to the following taxonomy [44]: *matrix factorization models*, *geometric models*, and *deep learning models*.

4.1 Matrix Factorization Models

RESCAL [41] is a approach that represents entities as vectors (h and t) and relations r as matrices.

DistMult [57] is a simplification of RESCAL for which the matrices of relations W_r have the constraint of being diagonal. This constraint lightens the model, because it considerably reduces the space of the parameters to be learned, however, becoming less expressive. For DistMult, all the relations are represented by a diagonal matrix, and are therefore considered as symmetrical.

ComplEx [52] is an extension of DistMult which represents relations and entities in a complex space ($h, r, t \in \mathbb{C}^d$). Thus, despite the same diagonal constraint as DistMult, it is possible to represent asymmetric relations thanks to the non-commutativity of the Hadamard product in the complex space.

QuatE [58] is an extension from ComplEx that goes beyond the complex-space to represent entities using quaternion embeddings and relations as rotations in the quaternion space. This models aims to offer better geometrical interpretations. QuatE is able to model symmetry, anti-symmetry and inversion.

SimplE [34] takes a 1927 approach [30] to tensor factorization (canonical polyadic decomposition) and adapts it to link prediction. This approach learns two independent vectors for each entity, one for the subject role and the other for the object role. They propose to make the representation of the subject and object vectors of one entity by relying on any relation r and its inverse r'.

TuckER [8] is a linear model for which a tensor of order 3 $T \in \mathbb{R}^{I \times J \times K}$ can be decomposed into a set of three matrices A, B and C and a core (a tensor of lower rank).

HolE [40] is an approach based on holographic embeddings. These make use of the circular correlation operator to compute interactions between latent features of entities and relations. That allows for a compositional representation of the relational data of the knowledge base.

4.2 Geometric Models

Structured Embedding [16] approach represents each relation by two matrices $M_r^h, M_r^t \in \mathbb{R}^{d \times d}$ that allow to perform projections specific to each relation of subject t and object t entities. This model thus makes it possible to distinguish the different types of relations, as in the role of subject and object of an entity.

TransE [15] utilized word embeddings and their capacity to account for the relations between words through translation operations between their vectors: $h+r \approx t$. Thus, $f(h,r,t) = -||h+r-t||_p$ where $p \in \{1,2\}$ is a hyper-parameter. TransE is limited regarding symmetrical or transitive relations, as well as for high cardinalities ($1 \dots N$ to 1, or 1 to $1 \dots N$ relations). This method has become popular because of its calculation efficiency.

TransH [56] addresses the expressivity limits of TransE for relations with cardinalities greater than 1. Each relation is represented as a hyperplane.

TransR [38] is an extension of the previously presented geometric models, which explicitly considers entities and relations as different objects, representing them in different vector spaces.

TransD [32] is another extension of TransR. Entities and relations are also represented in different vector spaces. The difference concerns the projection matrix which, unlike TransR, is not the same for all entities and only depends on the relation.

CrossE [59] extends the traditionnal models as it explicitly tackles crossover interactions, the bi-directional effects between entities and relations.

RotatE [50] represents relations as rotations between subject and object in complex space.

MurE [7] is the Euclidean counterpart of MuRP, a hyperbolic interaction model developed to effectively model hierarchies in KG.

KG2E [29] aims to model the uncertainties linked to entities and relations, compared to the number of observed triples containing these entities and relations. Thus, entities and relations are represented by distributions, more particularly Gaussian multivariate distributions.

4.3 Deep Learning Models

ConvE [22] is a multi-layer convolutional network model for link prediction, yielding the same performance as DistMult. This approach is particularly effective at modelling nodes with high in degree – which are common in highly-connected, complex knowledge graphs such as YAGO [49], DBpedia [6] or Freebase [14].

ERMLP [23] is a multi-layer perceptron based approach that uses a single hidden layer and represents entities and relations as vectors.

ProjE [47] is a neural network-based approach with a combination and a projection layer for candidate-entities.

5 Experimental Setup

The KGE methods presented in Section 4 allow more efficient use of knowledge bases by transforming them into vector space while maintaining their latent semantic properties. We present in the following section our experimental setup regarding the exploitation of these methods in the context of our dataset and additional data enrichment.

5.1 Evaluation

The evaluation is carried out by comparing the score of the actual triples against the score of all the other triples that could be predicted. Ideally, the score of the original triple is expected to have a better score against the others. This classification can be done according to two configurations: *raw* and *filtered*. In the *raw* configuration, all triples count for ranking, even valid triples belonging to the graph, while in the *filtered* configuration, these are not taken into account. We perform all the evaluations in a *filtered* configuration. From these ranks, the global metrics commonly used [44] are the following:

- **Mean Rank (MR)**: the average of these ranks. The lower this is, the more the model is able to predict the correct triples. Since this is an average, this measurement is very sensitive to outliers.
- **Mean Reciprocal Rank (MRR)**: the average of the inverse of the ranks. This metric is less sensitive to extreme values and is more common.
- **Hits@K**: the rate of predictions for which the rank is less than or equal to threshold K.

5.2 Data Pre-processing

The dataset and the ontology [5, 3] presented in Section 2 serve as the starting point for our study. First, we are interested in the instances that populate the ontology. Table 1 presents the object properties present in the ontology, along with their domain and their image (the classes between which they express a relation). For creating the sets for training, testing and validation, we eliminated the reverse object properties of the ontology to ensure the elimination of the risk of *data leakage*: the inverse relations could serve as information for the model during the testing stage. For example, determining if the triple (*Carrefour City*, has_address, "48 impasse du Ramier des Catalans") is true can be assisted by the realisation that has_address and is_address_of are inverse relations and by the triple (*"48 impasse du Ramier des Catalans", is_address_of, Carrefour City*).

5.3 Data Enrichment: External Verification

In order to allow an efficient verification of the information coming from the receipts, we are also interested in a data enrichment. We have about 15,000 triples (subject, predicate, object). This knowledge base, built solely from information from sales receipts, can be enriched using external resources. The reference base is intended to be easy to build and set up, based on existing databases or resources as structured as possible, such as company catalogs, in order to extract the price of products, the contact information of companies or even company registration information. To increase knowledge on companies, we utilized data from French national institute of statistics and economic studies (INSEE[6]). The INSEE provides freely available resources about companies, referred to as the SIRENE database.

[6] `https://www.insee.fr/en/accueil`

Table 1. Receipt ontology object properties.

Domain	Object Property	Reverse Property	Image
City	has_zipCode	is_zipCode_of	ZipCode
Company	has_contactDetail	is_contactDetail_of	ContactDetail
Company	has_adress	is_address_of	Address
Company	has_email_address	is_email_address_of	EmailAdress
Company	has_fax	is_fax_of	FaxNumber
Company	has_website	is_website_of	Website
Company	has_phone_number	is_phone_number_of	PhoneNumber
Company	issued	is_issued_by	Receipt
Product	has_expansion	is_expansion_of	Expansion
Company	has_registration	is_registration_of	Registration
Receipt	has_intermediate_payment		IntermediatePayment
Receipt	concerns_purchase		Product
Receipt	contains	is_written_on	Product, Registration, ContactDetail
SIREN	includes	is_component_of	SIRET, RCS, TVA IntraCommunity
City, ZipCode	part_of		Address
Company	is_located_at		City
Company	sells	is_sold_by	Product

Enrichment of Data from the SIRENE Database The methodology used, which aims to be fast and to build on existing resources, can be extended to other types of documents that contain the same relations (such as *has_address*, *has_SIRET*, etc.). Thus, we wish to incorporate data from the SIRENE database, which provides a database of data on French companies[7]. Given that the SIRENE registry has entries for 31 million French companies and that in the receipts corpus there are 387 different companies, we propose to measure the impact of the increase in data on the performance of fact-checking incrementally. To the 15,000 triples from the receipts, we added 1,000 to 9,000 triples from the SIRENE database, in steps of 1,000. This leaves us with ten different data sets, which size is presented in Table 2. The relations of the receipt ontology available in the SIRENE database are *has_zip_code*, *has_address*, *has_registration*, *is_component_of*, *part_of* and *is_located_at*. All of them relate to either registration information about the companies, or their addresses. In order to match the information from the receipt to the external reference database, certain approximations have been made. First, we consider that *City* class of ontology is equivalent to *Municipality*, field of the SIRENE database. We consider that in the receipt corpus we are dealing mainly with cities, a case in which this equivalence is respected. In addition, we chose to keep the relation *has_registration* as well as its hyponymous relations (*has_siren*, *has_siret* and *has_nic*). The relations that were added despite not being defined in the ontology are the following: *has_nic* was added in the same way as *has_siren* and *has_siret*; *has_main_activity* and *is_headquarter*. Table 2 compares the size of the obtained knowledge bases.

Address Alignment One of the limits of considering the textual content of documents as a knowledge graph is that lexical variation is not dealt with the

[7] `http://sirene.fr/siren/public/home`

Table 2. Size of the knowledge base extracted from sales receipts. The receipts_k with $k \in [1000, \ldots, 9000]$ represent the datasets resulting from the incremental data enrichment using the SIRENE database 5.3.

Dataset	No. Entity	No. Object Property	No. Triple
receipts	7,108	16	14,379
receipts_1000	7,613	21	15,379
receipts_2000	8,081	21	16,379
receipts_3000	8,542	21	17,379
receipts_4000	8,990	21	18,379
receipts_5000	9,406	21	19,379
receipts_6000	9,808	21	20,379
receipts_7000	10,196	21	21,379
receipts_8000	10,500	21	22,379
receipts_9000	10,918	21	23,379

extracted text from the receipts becomes the label of the entities. In the ontology, the addresses are broken down into their components (ZipCode and City). However, one address expressed in two different ways would not be recognized as the same. We therefore used the application programming interface (API) of the National Address Database (BAN[8]), the only database of addresses officially recognized by the French administration. We added a query allowing to associate each address with its id in the BAN. The id associated with each address depends on its specificity: 17300_7593_00033 is the id of *33 Rue de la Scierie 17000 La Rochelle*; 17300_7593 is the id of *Rue de la Scierie 17000 La Rochelle*. This approach makes it possible to reconcile identical addresses, but it also constitutes an implicit verification step. Addresses that were not associated with an id were kept with their full text.

6 Results

We trained the methods presented in Section 3 on the receipts corpus with and without the different levels of data enrichment presented in Table 5.3[9].

Results without Data Enrichment As we can see in Figure 3, the best results come from matrix factorization models, particularly QuatE. DistMult, ComplEx and QuatE, are both based on the RESCAL method. However, ComplEx shows very poor results, even lower than RESCAL. QuatE represents entites in the quaternion space, which extends the expressivity of the model compared to DistMult, which represents every relation with a diagonal matrix and imposes

[8] `https://api.gouv.fr/les-api/base-adresse-nationale`

[9] The dataset has been split into a training and test set (80% and 20% respectively) thanks to the PyKEEN library `https://github.com/pykeen/pykeen` [2], to avoid redundant triples being found both in training and test. The previously presented methods are implemented by PyKEEN, library that we chose to use for its completeness, flexibility and ease of use.

symmetry. Second best results come from geometric models, especially MuRE, TransD and RotatE. Lastly, deep learning models show very low performance: the small size of the receipt dataset can explain this low result. We can observe that the methods struggle to accurately represent the entities and relations found in the receipts.

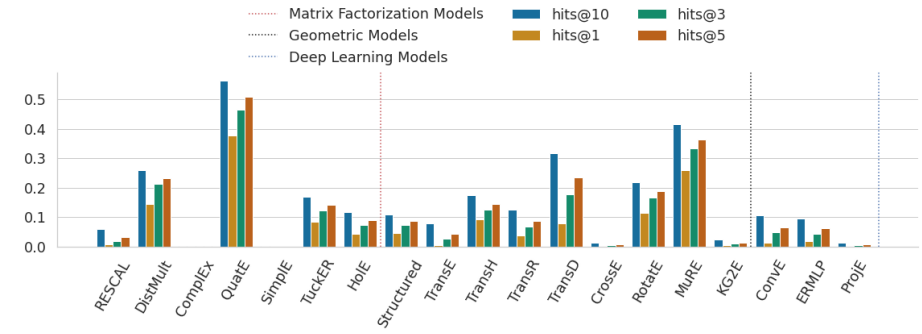

Fig. 2. Hits@K without data enrichment.

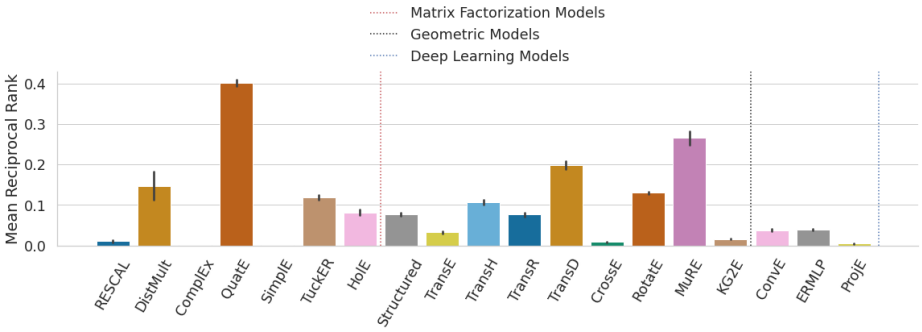

Fig. 3. MRR scores without data enrichment.

Figure 4 shows the receipt entities embeddings as represented by the different models. We can observe that most methods do not allow for an efficient receipt classification: all the documents are represented in a unique cluster. This is in accordance with the poor results the methods show in regards to the mean rank scores in Figure 2 and mean reciprocal rank scores in Figure 3. The only method that yields document classification results is QuatE, as it is able to distinguish receipts from others according to their issuing company. However, the

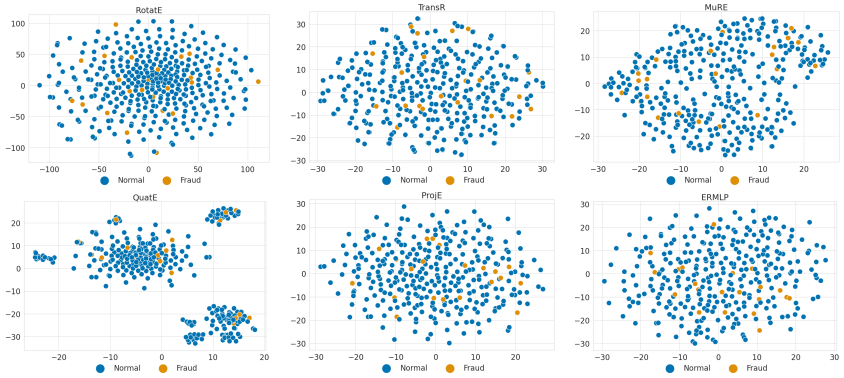

Fig. 4. Representation of the entities without data enrichment with T-SNE [53]. We only chose the models that obtained a rather higher performance in Figure 2

.

classification is not enough to detect forgery, as we can see with the distribution of forged receipts (in orange in Figure 4). When analyzing the MR results, the lower they are, the better, as it means the correct triples have a lower rank. The opposite goes for MRR, which is the inverse of the MR. The lowest results are for the SimplE and ComplEx methods (Figure 3), both matrix factorization methods. One of our data pre-processing stages was the elimination of inverse triples in order to avoid the risk of test leakage. However, SimplE relies on inverse triples to accurately represent the entities and relations of the KG: our implementation choice explains the rather weak results of the SimplE method.

We performed an additional evaluation by filtering the relations in the test set in order to understand what relations were learned better by QuatE (as seen in Table 3). We also provide the number of triples containing each relation in order to be cautious when interpreting those results.

Table 3. Hits@10 results of the QuatE model for the six best learned relations.

Relation	**Hits@10**	**Triple Count**
has_zipCode	66%	343
has_address	66%	243
has_phone_number	61%	315
sells	43%	4,555
contains	41%	1,906
concerns_purchase	39%	4,555

It can be noted in Table 3 that the best-learned relations are related to the contact information of the companies (zip code, address, and phone number). However, relations expressing the structure of the receipts (contains and concerns_purchase) have weaker results. Learning the semantic structure of the KB

built from the receipt dataset proves to be a difficult task. Most of the information available is related to the purchases made and which companies they were made from. That induces a bias, and as we are approaching the problem as a prediction task, this implies that we may be dealing with an underlying consumer behavior study that is not the object of our research, and for which we have no ground truth. Indeed, we are interested in learning the document structure and the ability of KGE to assess document coherence.

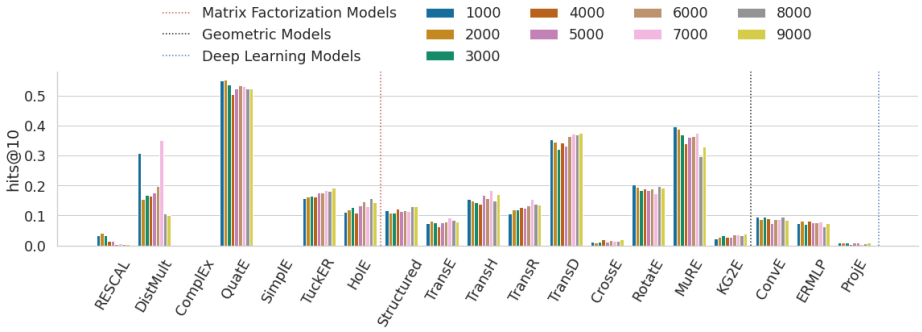

Fig. 5. Hits@10 scores variation across the different levels of data enrichment.

Results with Data Enrichment We then evaluated the models across the different data enrichment steps. As we trained the different models on different size datasets, we do not show the MR results, as they cannot be compared: with more triples to sort, an equivalent performance would rank the correct triples lower.

Figure 5 shows the variation of the hits@K scores for all the models with triples from the SIRENE database. QuatE shows the best performance across all enrichment steps. However, we can observe how the enrichment process improves the results of several methods (DistMult, MuRE, RotatE, TransD, HolE, and TransH). That improvement is however not linear but could be explained by the nature of the receipt dataset. Indeed, the collected receipts mainly come from the same city, and thus contain very local information, while the enrichment comes from a national database and was done randomly.

7 Discussion and Conclusions

Document forgery detection poses a challenge for KGE methods aiming to represent the information in a document to authenticate it. However, data augmentation can improve the results. A perspective of our work was to approach the data augmentation in a specialized manner. Instead of adding triples with

a relation semantic matching criteria, we could focus more on the documents to authenticate and add reference information about the companies, addresses, products, and other entities that are more prevalent in the data.

Another perspective would be to turn to methods that take into account literal information [27]. Most KGE methods focus on the entities and the relations between them, without taking into account the additional information available. In the particular case of forgery detection, a lot of relevant information is ignored (prices, dates, quantities, etc.).

A limitation concerning the data used to train and evaluate KGEs [45] is their imbalanced nature: 15% of the entities of FB15K is contained in 80% of the triples. For example, the entity "United States" is contained in almost all nationality relations. It is more profitable in terms of performance to predict an American nationality regardless of the subject entity than to learn this relationship with its underlying structure. Our dataset suffers from this bias.

The evaluation metrics are holistic as they do not distinguish the relations by type [45]. This holistic evaluation presents several shortcomings in the context of fraud detection, as the different relations play different roles in the semantic coherence structure of the document. Some relations are harder to predict because of their structural features, in particular, the high cardinality of the relations related to the products sold in the receipts - having a higher number of target peers for a triple makes link prediction harder [44]. Those relations also play a less important role in document authentication, as they are more related to user behavior than information we have the means to verify.

Our study, thus, confronted the scientific obstacles linked to the opacity of fact-checking methods for document fraud detection based on knowledge bases. Experimenting with these different methods in the concrete and complex application case of fraud detection however made it possible to set up an extrinsic evaluation of these methods.

References

1. Abiteboul, S.: Semistructured data: from practice to theory. In: Proceedings 16th Annual IEEE Symposium on Logic in Computer Science. IEEE (2001)
2. Ali, M., Berrendorf, M., Hoyt, C.T., Vermue, L., Sharifzadeh, S., Tresp, V., Lehmann, J.: Pykeen 1.0: A python library for training and evaluating knowledge graph emebddings (2020)
3. Artaud, C., Doucet, A., Ogier, J.M., d'Andecy, V.P.: Receipt dataset for fraud detection. In: First International Workshop on Computational Document Forensics (2017)
4. Artaud, C., Sidère, N., Doucet, A., Ogier, J.M., Yooz, V.P.D.: Find it! fraud detection contest report. In: 2018 24th International Conference on Pattern Recognition (ICPR). IEEE (2018)
5. Artaud, C.: Détection des fraudes : de l'image à la sémantique du contenu : application à la vérification des informations extraites d'un corpus de tickets de caisse. PhD Thesis (2019)
6. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data. In: Aberer, K., Choi, K.S., Noy, N., Allemang, D.,

Lee, K.I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) The Semantic Web. pp. 722–735. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
7. Balazevic, I., Allen, C., Hospedales, T.: Multi-relational poincaré graph embeddings. Advances in Neural Information Processing Systems **32** (2019)
8. Balažević, I., Allen, C., Hospedales, T.M.: Tucker: Tensor factorization for knowledge graph completion (2019)
9. Barzilay, R., Lapata, M.: Modeling local coherence: An entity-based approach. Computational Linguistics **34**(1) (2008)
10. Behera, T.K., Panigrahi, S.: Credit card fraud detection: a hybrid approach using fuzzy clustering & neural network. In: 2015 Second International Conference on Advances in Computing and Communication Engineering. IEEE (2015)
11. Berti-Équille, L., Borge-Holthoefer, J.: Veracity of Data: From Truth Discovery Computation Algorithms to Models of Misinformation Dynamics. Synthesis Lectures on Data Management **7**(3), 1–155 (Dec 2015)
12. Bertrand, R., Gomez-Kramer, P., Terrades, O.R., Franco, P., Ogier, J.M.: A System Based on Intrinsic Features for Fraudulent Document Detection. In: 2013 12th International Conference on Document Analysis and Recognition. pp. 106–110. IEEE, Washington, DC, USA (Aug 2013)
13. Bertrand, R., Terrades, O.R., Gomez-Krämer, P., Franco, P., Ogier, J.M.: A conditional random field model for font forgery detection. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). IEEE (2015)
14. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: A collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. p. 1247–1250. SIGMOD '08, Association for Computing Machinery, New York, NY, USA (2008)
15. Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., Yakhnenko, O.: Translating Embeddings for Modeling Multi-Relational Data. In: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2. pp. 2787–2795. NIPS'13, Curran Associates Inc., Red Hook, NY, USA (2013), event-place: Lake Tahoe, Nevada
16. Bordes, A., Weston, J., Collobert, R., Bengio, Y.: Learning Structured Embeddings of Knowledge Bases. In: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence. pp. 301–306. AAAI'11, AAAI Press (2011), event-place: San Francisco, California
17. Cozzolino, D., Gragnaniello, D., Verdoliva, L.: Image forgery detection through residual-based local descriptors and block-matching. In: 2014 IEEE international conference on image processing (ICIP). IEEE (2014)
18. Cozzolino, D., Poggi, G., Verdoliva, L.: Efficient dense-field copy–move forgery detection. IEEE Transactions on Information Forensics and Security **10**(11) (2015)
19. Cozzolino, D., Verdoliva, L.: Camera-based image forgery localization using convolutional neural networks. In: 2018 26th European Signal Processing Conference (EUSIPCO). IEEE (2018)
20. Cozzolino, D., Verdoliva, L.: Noiseprint: A cnn-based camera model fingerprint (2018)
21. Cruz, F., Sidere, N., Coustaty, M., d'Andecy, V.P., Ogier, J.M.: Local binary patterns for document forgery detection. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1. IEEE (2017)
22. Dettmers, T., Minervini, P., Stenetorp, P., Riedel, S.: Convolutional 2d knowledge graph embeddings (2018)

23. Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., Zhang, W.: Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 601–610. KDD '14, Association for Computing Machinery, New York, NY, USA (2014), event-place: New York, New York, USA
24. EulerHermes-DFCG: Plus de 7 entreprises sur 10 ont subi au moins une tentative de fraude cette année `https://www.eulerhermes.fr/actualites/etude-fraude-2020.html`
25. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Transactions on Information Forensics and Security **7**(3) (2012)
26. Galárraga, L., Teflioudi, C., Hose, K., Suchanek, F.: Fast Rule Mining in Ontological Knowledge Bases with AMIE+. The VLDB Journal (2015), publisher: Springer
27. Gesese, G.A., Biswas, R., Alam, M., Sack, H.: A Survey on Knowledge Graph Embeddings with Literals: Which model links better Literal-ly? (May 2020)
28. Goyal, N., Sachdeva, N., Kumaraguru, P.: Spy the lie: Fraudulent jobs detection in recruitment domain using knowledge graphs. In: International Conference on Knowledge Science, Engineering and Management. Springer (2021)
29. He, S., Liu, K., Ji, G., Zhao, J.: Learning to represent knowledge graphs with gaussian embedding. In: Proceedings of the 24th ACM international on conference on information and knowledge management (2015)
30. Hitchcock, F.L.: The expression of a tensor or a polyadic as a sum of products. Journal of Mathematics and Physics **6**(1-4) (1927)
31. Huynh, V.P., Papotti, P.: A benchmark for fact checking algorithms built on knowledge bases. In: CIKM 2019, 28th ACM International Conference on Information and Knowledge Management, November 3rd-7th, 2019, Beijing, China. Beijing, CHINE (2019)
32. Ji, G., He, S., Xu, L., Liu, K., Zhao, J.: Knowledge graph embedding via dynamic mapping matrix. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers) (2015)
33. Ji, S., Pan, S., Cambria, E., Marttinen, P., Yu, P.S.: A Survey on Knowledge Graphs: Representation, Acquisition and Applications (Aug 2020)
34. Kazemi, S.M., Poole, D.: Simple embedding for link prediction in knowledge graphs (2018)
35. Kim, J., Kim, H.J., Kim, H.: Fraud detection for job placement using hierarchical clusters-based deep neural networks. Applied Intelligence **49**(8) (2019)
36. Kowshalya, G., Nandhini, M.: Predicting fraudulent claims in automobile insurance. In: 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT). IEEE (2018)
37. Li, Y., Yan, C., Liu, W., Li, M.: Research and application of random forest model in mining automobile insurance fraud. In: 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE (2016)
38. Lin, Y., Liu, Z., Sun, M., Liu, Y., Zhu, X.: Learning entity and relation embeddings for knowledge graph completion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 29 (2015)
39. Mishra, A., Ghorpade, C.: Credit card fraud detection on the skewed data using various classification and ensemble techniques. In: 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS). IEEE (2018)

40. Nickel, M., Rosasco, L., Poggio, T.: Holographic embeddings of knowledge graphs. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 30 (2016)
41. Nickel, M., Tresp, V., Kriegel, H.P.: A Three-Way Model for Collective Learning on Multi-Relational Data. In: Proceedings of the 28th International Conference on Machine Learning, ICML 2011. pp. 809–816 (2011)
42. Rabah, C.B., Coatrieux, G., Abdelfattah, R.: The supatlantique scanned documents database for digital image forensics purposes. In: 2020 IEEE International Conference on Image Processing (ICIP). IEEE (2020)
43. Rizki, A.A., Surjandari, I., Wayasti, R.A.: Data mining application to detect financial fraud in indonesia's public companies. In: 2017 3rd International Conference on Science in Information Technology (ICSITech). IEEE (2017)
44. Rossi, A., Firmani, D., Matinata, A., Merialdo, P., Barbosa, D.: Knowledge Graph Embedding for Link Prediction: A Comparative Analysis (Mar 2020)
45. Rossi, A., Matinata, A.: Knowledge graph embeddings: Are relation-learning models learning relations? In: EDBT/ICDT Workshops (2020)
46. Shen, A., Mistica, M., Salehi, B., Li, H., Baldwin, T., Qi, J.: Evaluating document coherence modeling. Transactions of the Association for Computational Linguistics **9**, 621–640 (07 2021)
47. Shi, B., Weninger, T.: Proje: Embedding projection for knowledge graph completion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
48. Sidere, N., Cruz, F., Coustaty, M., Ogier, J.M.: A dataset for forgery detection and spotting in document images. In: 2017 Seventh International Conference on Emerging Security Technologies (EST). IEEE (2017)
49. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: A core of semantic knowledge. In: Proceedings of the 16th International Conference on World Wide Web. p. 697–706. WWW '07, Association for Computing Machinery, New York, NY, USA (2007)
50. Sun, Z., Deng, Z.H., Nie, J.Y., Tang, J.: Rotate: Knowledge graph embedding by relational rotation in complex space (2019)
51. Thorne, J., Vlachos, A.: Automated Fact Checking: Task formulations, methods and future directions. CoRR (2018)
52. Trouillon, T., Welbl, J., Riedel, S., Éric Gaussier, Bouchard, G.: Complex embeddings for simple link prediction (2016)
53. Van Der Maaten, L.: Accelerating t-sne using tree-based algorithms. The Journal of Machine Learning Research **15**(1) (2014)
54. Vidros, S., Kolias, C., Kambourakis, G., Akoglu, L.: Automatic detection of online recruitment frauds: Characteristics, methods, and a public dataset. Future Internet **9**(1) (2017)
55. Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge Graph Embedding: A Survey of Approaches and Applications. IEEE Transactions on Knowledge and Data Engineering **29**(12), 2724–2743 (2017)
56. Wang, Z., Zhang, J., Feng, J., Chen, Z.: Knowledge graph embedding by translating on hyperplanes. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 28 (2014)
57. Yang, B., tau Yih, W., He, X., Gao, J., Deng, L.: Embedding entities and relations for learning and inference in knowledge bases (2015)
58. Zhang, S., Tay, Y., Yao, L., Liu, Q.: Quaternion knowledge graph embeddings (2019)
59. Zhang, W., Paudel, B., Zhang, W., Bernstein, A., Chen, H.: Interaction embeddings for prediction and explanation in knowledge graphs. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (2019)