

Enhancement of Record Linkage by Using Attributes containing Natural Language Text

Philipp Behnen^a, Felix Kruse^a and Jorge Marx Gómez^a

^aUniversität Oldenburg, Department VLBA, Ammerländer Heerstraße 114-118, 26129 Oldenburg, Lower Saxony, Germany

Abstract

The integration of external and internal data sources is becoming increasingly important, as decision-makers depend on complete information. Often the data sources to be integrated do not have a common and unique identifier. In these cases, the data sources must be integrated by comparing the available common attributes of the entity, the so-called record linkage. There are similarity measures for attributes that contain strings like the company name or numbers such as turnover. Attributes that contain natural language text, such as company descriptions, are still unused. This research paper describes a research project on using natural language text attributes applying Machine Learning for entity matching. The use of natural language text attributes is intended to improve the results of entity matching and, thus, data integration.

Keywords

data integration, record linkage, natural language processing, word embedding, company descriptions

1. Introduction

Internal and external data sources are crucial for supporting decision-making processes in research and industry, as they can contain relevant information for the decision-maker [1]. The relevant information for the particular decision is rarely available in one data source. Therefore, commonly more than one data source is needed. These data sources contain different complementary or identical information that the decision-maker needs [2]. Firstly, the different data sources must be integrated to make this information base accessible to the decision-maker. In the best case, a unique identification number for the entities among the data sources to be integrated is given. If there is no unique identification number for the entities of the data sources, the entity's existing attributes must be used for a similarity measurement to integrate them. These procedures are defined by the terms entity matching (EM) or record linkage (RL). These procedures are used to identify which data records belong to the same real-world entity. For example, a company may be represented by a name, the address, and a description (cf. Table 1). All available attributes of the entity should be used to compare the records to perform successful data integration using RL. String similarity measures already exist for comparing textual

In A. Martin, K. Hinkelmann, H.-G. Fill, A. Gerber, D. Lenat, R. Stolle, F. van Harmelen (Eds.), Proceedings of the AAAI 2021 Spring Symposium on Combining Machine Learning and Knowledge Engineering (AAAI-MAKE 2021) - Stanford University, Palo Alto, California, USA, March 22-24, 2021.

✉ philipp.behnen@uol.de (P. Behnen); felix.kruse@uol.de (F. Kruse); jorge.marx.gomez@uol.de (J.M. Gómez)

ORCID 0000-0000-0000-0000 (P. Behnen); 0000-0000-0000-0000 (F. Kruse); 0000-0000-0000-0000 (J.M. Gómez)

© 2021 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).


 CEUR Workshop Proceedings (CEUR-WS.org)

Table 1

Example of attributes that may be used in RL

Field	Record 1	Record 2	Record 3
Name	Amazon Corp.	AMZ	VW AG
Address	Example Street 123	Null	Wolfsburg, Germany
Desc.	Amazon is a Seattle based e-commerce platform for consumers and sellers delivering products all around the world	Online marketplace for all kinds of goods from the US west coast, with a two day delivery service called Prime.	car manufacturer located at lower saxony in Germany acting globally that exists since 1937

Table 2

Similarity score of different unsupervised matching approaches on sample company descriptions

ID	Description	ID	Description	Jaro-Wink.	TF-IDF (Cos.)	BERT (Cos.)	Match
1	Amazon is a Seattle based e-commerce platform for consumers and sellers delivering products all around the world	2	Online marketplace for all kinds of goods from the US west coast, with a two day delivery service called Prime.	84%	10%	86%	Yes
1	Amazon is a Seattle based e-commerce platform for consumers and sellers delivering products all around the world	3	car manufacturer located at lower saxony in Germany acting globally that exists since 1937	88%	3%	25%	No

attributes such as name and address. Edit distances like Levenshtein or Jaro-Winkler can detect simple changes of characters. Token-based methods like Jaccard and Tf-IDF can detect token permutations. Hybrid methods such as Monge-Elkan and Soft-TF-IDF can detect token permutations with character changes [3]. These traditional similarity methods reach their limits when it comes to recognizing semantic heterogeneities such as acronyms and abbreviations. For example, no classic string similarity measure recognizes that "AAAI-MAKE 2021" and "Association for the Advancement of AI Spring Symposium" refer to the same conference. To cope with the semantic heterogeneity and to use attributes with natural language texts (such as the description in Table 1) for a similarity comparison, methods from the field of "Natural Language Processing" (NLP) can be used [3]. One of these methods are word embeddings, which have been established as the standard solution for different NLP tasks [4]. Word embeddings are machine learning methods trained to represent texts or words as numbers (vectors) so that they can be compared using similarity measures such as the cosine similarity [5]. These models can recognize language patterns and process syntactic and semantic heterogeneities such as sentence structure, grammar, and negations. In the past two years, especially modifications of the "Bidirectional Encoder Representations from Transformers" (BERT) architecture have become state-of-the-art in various NLP fields when pre-trained models are used as a basis for

fine-tuning on specific downstream tasks [6]. Word embeddings like BERT can be used in RL to include attributes such as company descriptions in the similarity determination [4]. This paper describes a research project that aims to investigate which methods can be used to make attributes with natural language texts such as company descriptions usable for RL. Table 2 shows that the existing string similarity measures cannot be applied directly to the company description because the correct tuples cannot be selected. With this paper we make the following research contribution:

- We present the current state of the art of research on the use of natural language text attributes in RL
- We describe our approach to building a dataset to develop and evaluate suitable methods for natural language text attributes
- We show conceptual methods to solve the RL problem with natural language text attributes

Therefore, this paper is structured as follows: In section 2, the specific problem statement is addressed and related work is collected and analyzed by performing a literature review. The shortcomings of the related work for the use case of our work are addressed in section 3, where a natural language text based evaluation set for RL is created and the process is described. In section 4, we propose different natural language text based RL approaches and provide the performance of baseline models on our dataset as preliminary results of our ongoing work. In section 5 future work is discussed and a conclusion is drawn.

2. Problem statement and related work

Since the company descriptions are natural language texts, different NLP approaches could be used. For example, Named Entity Recognition (NER) could be used to extract further descriptive attributes for the entity of the texts, which could be compared by classical string similarity measures. Word embeddings could convert the texts directly into vectors and compare them for similarity using measures like the cosine similarity. Another problem that NLP approaches can solve is when the company descriptions are available in different languages. Word embeddings like BERT may learn language representations on multiple languages at once [7]. There exists research on RL and the use of attributes containing natural language texts. With the help of a qualitative literature analysis [8] and based on the literature review by Kruse et al. [9], relevant papers were identified. The search strategy of the qualitative literature analysis is methodically based on common standards for literature reviews (see Webster and Watson [8]), which searches for articles by keywords for relevant topics rather than by specific authors. The search query is the same for each database and is applied to title, abstract and keywords. Publications since 2017 are evaluated. Specifically, the following query was made:

duplicate detection OR record matching OR entity matching OR entity
resolution OR record linkage OR entity linking

Table 3

Amount of publications found in literature review

Database	Found	First screening	Full text screening
ACM DL	225	24	9
arXiv	231	6	1
IEEE	232	15	6
ScienceDirect	245	4	0
FW & BW-Search	/	/	2
Total			18

The first screening consists of the analysis of title, abstract and conclusion and serves as a preliminary selection of relevant papers. This is followed by a forward-backward search, identifying the cited contributions in these papers as well as publications that refer to these papers [8]. The final step is the full-text screening, where a final selection of relevant papers is made. Relevant publications apply techniques to perform RL in practice. Text Matching without focus on entities (e.g. plagiarism detection, comparing message texts) may be important for the later solution finding, but is not considered relevant for the classification of this work in existing research by the literature review. The same applies to publications using techniques that directly compare texts with an entity name from a knowledge base (Entity Linking). The result of the search is shown in Table 3. Based on the literature review of [9] and our work, theory-based inductive categories [27] were formed. Identified contributions are distinguished with respect to the entities to be compared (e.g. companies, products or persons), since the domain of company data is particularly relevant for our work. The same applies to the indication whether existing publications have tested multilingual approaches. (Word-) Embeddings provide state-of-the-art results in various application areas and are suitable for use in the RL process [4], therefore the used procedures represent a further category. According to [6] additional steps like *Fine tuning* and *Transfer learning* may lead to improvements compared to direct supervised learning on a training data set. Whether such techniques are used is another category. After about 50 % of the full text screening, categories were refined or added based on this. It was found that some of the papers perform RL on texts (see [18] and [20]), which for this paper can be called semi-structured attributes and are similar in structure to product titles. Natural language texts are, according to the definition in this work, complete sentences or continuous texts. In the course of the deductive category formation according to [27] the categories *semi-structured text* and *natural-language text* were included in the classification scheme. The developed concept matrix according to [8] is shown in Table 4. In the following, the contributions are described in more detail with regard to the methods or data sets used, thus explaining the reasons for their classification in the concept matrix.

Mudgal et al. use different techniques to use both structured attributes and natural language texts in RL [4]. For the processing of texts from the areas of product and company descriptions, the Word Embeddings *GloVe* and *FastText* are used, among others. A Python module based on *FastText* is provided under the name *DeepMatcher*. The authors see finetuning as a possible improvement of their work. Li et al. partially fills this research gap and uses a pre-trained BERT

Table 4
Comparison to relevant work

Publication	Semi-structured text	Natural language text	Company entities	Fine-tuning	Transfer Learning	Multilingual approaches	Embedding Architecture
Ebraheem et al. [10]	X	X				(X)	GloVe/Word2Vec
Li et al. [11]	X						
Mudgal et al. [4]	X	X	X				FastText
Ristoski et al. [12]	X	X	X				paragraph2vec
Schneider et al. [13]	X	X					Word2Vec
Sim and Borthwick [14]	X						Record2Vec
Song et al. [15]	X						
Thirumuruganathan et al. [16]	X			X	X		FastText
Brunner and Stockinger [17]	X						
Gschwind et al. [18]	X		X				
Javdani et al. [19]	X						Word2Vec
Nie et al. [20]	X						FastText
Primpeli et al. [21]	X						
Zhao and He [22]	X			X	X		
Li et al. [23]	X	X	X	X			BERT
Meduri et al. [24]	X						
Wu et al. [25]	X		X				
Zhang et al. [26]	X	X	X		X		FastText
Our Work		X	X	X	X	X	Different Transformers

architecture to perform RL[23].

In the papers [11, 24], RL is carried out with product titles. The product titles not only contain atomic attributes but also consist of natural language text elements. However, company descriptions vary in their format to product titles because they consist of syntactically correct sentences. It would be interesting to research whether the same approaches may work in both cases and domains. We identified five papers using company descriptions [12, 18, 23, 25, 26]. Each one of these papers uses the Deepmatcher data set [4] for their experiments. The Abt-Company set contained in Deepmatcher consists of the first paragraphs of Wikipedia articles and texts from company websites. Data from commercial databases, such as Crunchbase, are not prevalent. Only one of the papers uses state-of-the-art BERT-Embeddings (or other Transformer architectures) for their experiments [23] but does not incorporate other attributes. The performance of an RL system that only uses the description was not measured. Regarding multilingual RL, we have not found any publications performing experiments, which also provides research opportunities. Additionally, no paper attempts to match company descriptions

with the help of word or sentence representations. Although, unsupervised matching using pre-trained BERT vectors and cosine similarity has shown promising results in our first experiments, which will be presented in section 4.

3. Creating a natural language text based evaluation set for RL

For the training of models and the evaluation annotated text pairs are needed which represent matches or non-matches. The goal is to efficiently create as many data sets as possible, ideally with little manual effort. For the purpose of our research a purely German-language dataset as well as a mixed language dataset (German + English) is needed. To make the models trained for the RL task robust and flexible the texts should vary in length.

Existing datasets like the Abt-Company dataset contained in the Deepmatcher framework are not able to suffice for our work for various reasons. Firstly, the dataset does not inherently contain texts in languages other than English. Secondly, the data sources used are Wikipedia texts and crawled company websites, and the latter appear to vary in data quality and which ultimately might affect performance. Thirdly, no enterprise databases, which are regarded highly relevant for the practitioners in our research project, are included in the data. Since linking records from professional data sources is one key aspect of our use case driven project, training data from such sources might prove helpful to find the best approaches to solve the RL task.

Based on our requirements and building on the shortcomings of the existing datasets, we propose a evaluation set for company description based RL, primarily for German and English texts. However, the proposed approach is easily adaptable to create evaluation sets for various languages.

We have identified three real-world data sources that are used to build the evaluation dataset. The first data source is Wikidata¹. From the Wikidata 258,109 companies were extracted, that are represented via a unique identifier. The Wikipedia API² provides functions to query individual components of a Wikipedia page using such an wikidata identifier. For all the entities available, we used the API to query the English company descriptions from the Wikipedia. The contained text is available as readable raw text and does not contain HTML or Wikimedia tags (markup language), which makes later processing easier and generally can be regarded as high quality data.

The Wikipedia API may also be used to determine whether a Wikipedia page exists in other languages and what the corresponding Wikipedia link is³. For a total of 20,126 entities, German-language pages are also available. These 20,126 entities were selected for further consideration in our work, since different texts for the same entity are present (English as well as German) and thus may directly be used to train or evaluate (cross-lingual) textual RL models without the need of manual annotation. The second data source used is the English-language company database Crunchbase⁴. Crunchbase offers two data sets containing companies with no common

¹<https://www.wikidata.org/>

²https://www.mediawiki.org/wiki/API:Main_page/de

³e.g. via <https://en.Wikipedia.org/w/api.php?format=json&action=query&titles=Lufthansa&prop=langlinks>

⁴<https://www.crunchbase.com/home>

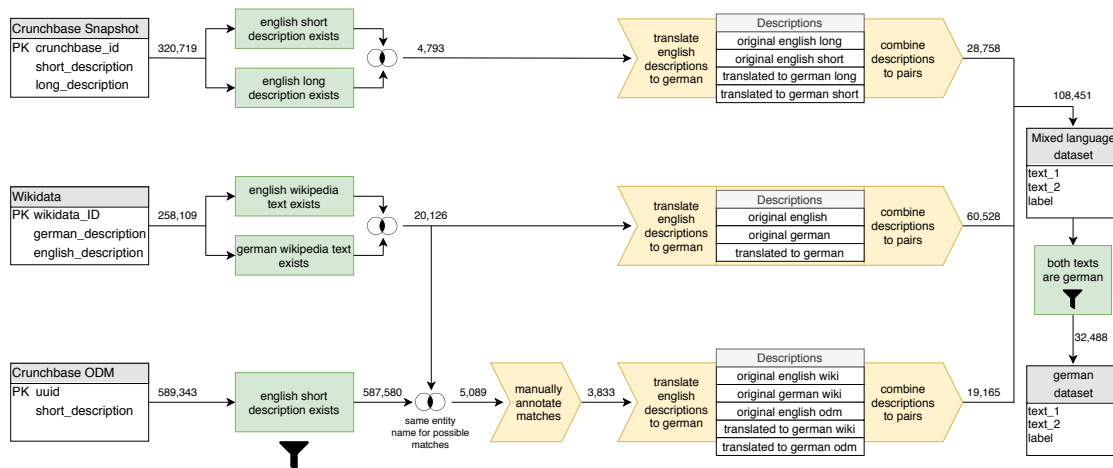


Figure 1: Process of creating the evaluation set with the number of matches (rows) in selected steps

identifier. The Crunchbase Open Data Map⁵ (ODM) and the Crunchbase Snapshot 2013⁶. The Crunchbase Snapshot 2013 contains 320,719 company records with an English short and long description (*short_description* and *overview*). 4,793 records contain a short and a long description and can be used as matches for the evaluation dataset. The Crunchbase ODM contains 589,343 company records with an English short description (*short_description*), but no alternative description. We manually labeled the Wikidata and Crunchbase ODM datasets to get more match samples between company descriptions. For this, we extracted all record pairs that have names that match exactly, in the hope of finding match candidates with a high probability of actually being matches. Of the 5,089 pairs found, we have labeled 3833 pairs as actual matches. From the three data sources, only Wikidata contains German texts. For the creation of the evaluation set with exclusively German descriptions, every available English text is translated using a machine learning model. For this purpose a pre-trained transformer from the Python library Huggingface/Transformers [28] is used. The model MarianMT⁷ was chosen, because it delivered an acceptable calculation time and quality in own tests. Minor grammar mistakes are no problem as long as the meaning of the text is not changed. The quality of the translated texts is considered sufficient for the application. By translating the English Wikipedia texts, the Crunchbase ODM short description and the CrunchbaseSnapshot short and long descriptions there are exactly two German descriptions for each entity of the three tables available for training and evaluating models. After collecting, labelling and translating, the following amount of matches and the available attributes are:

- 20,126 entities from Wikipedia with original English, original German and translated German descriptions

⁵Powered by Crunchbase: <https://data.crunchbase.com/docs/open-data-map>

⁶Crunchbase 2013 Snapshot © 2013, Creative Commons Attribution License [CC-BY], <https://data.crunchbase.com/docs/2013-snapshot>

⁷Precisely: Marian MT Helsinki-NLP/opus-en-de

Table 5

Extract of the resulting mixed language evaluation set

Text1	Text2	Match
Pixelpipe is a media distribution gateway allowing users to publish text, photos, video, audio and documents.	(name removed) is a web gateway that allows mobile desktop and server applications to publish content (photo, video, audio, text, file) and have it distributed out to social networks, websites and blogs around the world.	1
Tesla Motors accelerates the transition to electric mobility with a range of increasingly affordable electric cars.	TESLA (named after Nikola Tesla, later explained as abbreviation from technika slaboprouda, meaning "low-voltage technology") was a state-owned electrotechnical conglomerate in the former Czechoslovakia.	0

- 4,793 entities from crunchbase snapshot with original English short description, original English long description, translated German short descriptions and translated German long descriptions
- 3,833 entities from manually annotating matches between crunchbase odm and Wikipedia with original English Wikipedia, original German Wikipedia, translated German Wikipedia, original English crunchbase and translated German crunchbase descriptions

Due to the different available description texts for entities within a data source, text pairs with matches can be created without a manual annotation process. For example, after collecting the Wikipedia data and translating it, one English and two German description texts are available for each entity. In combination with each other, this results in a total of 3 training examples (English + German1, English + German2, German1 + German2). For the cross-language evaluation set, all of those combinations can be used, whereas for the pure German data set only the text pairs with German descriptions can be used. Figure 1 shows the whole process of data sources or description texts that can be combined by the automated and manual annotation process with the number of matches found. The set of matches resulting from the combination is 32,488 for the purely German-language and 108,451 for the mixed language dataset. For negative samples, random texts of the set are chosen in a way that every entity in the dataset has the same amount of matches and non matches, equally distributing the classes among the dataset. This doubles the amount of samples in each dataset. While the texts extracted from the three data sources generally were preserved as is and were not changed, there is one notable exception: In 50% of the cases the name of the entity was removed from the respective description text, if a name was available. The goal of this measure is to force the developed models to abstract beyond the entity name and not to restrict themselves to this single property. This essentially leads to 25% of samples where it is guaranteed that neither descriptions contains the company name, 50% of samples where exactly one description has the company name removed (25% and 25% of samples that are completely unchanged).

For the German language as well as the mixed language dataset, 10% of the examples have been retained for validation, and models can be trained on the remaining 90% (if the learning process is supervised). From the manual annotation of possible matches between Crunchbase ODM and Wikipedia, 708 non-matches of entities with the same name were also identified. Due

to the same name, the matching of these entities poses a special challenge for the developed models and is only used in the validation data set. The goal of this procedure is to measure a degree of abstraction beyond the company name. Good models should be able to exclude a match despite the same name due to the remaining text information. An extract of the resulting evaluation set is shown in the Table 5.

4. Design space for textual RL and preliminary results

Our research’s overall goal is to use natural language text attributes for RL to get better match results. This will be developed for the entity company using the proposed evaluation set containing company descriptions from different data sources. Following up on the research done, two research questions have been defined to achieve this goal:

- What is the best method for matching company descriptions?
- How does the use of the company description affect the overall performance of integrating data sources with the entity company?

Laboratory experiments will be conducted to answer the first research question. In addition to the Abt-Company data set in the widely used Deepmatcher framework, company data sources such as Crunchbase and Wikidata will be part of the experiments. For this, we will use the proposed evaluation set and may add more data sources to obtain generally valid results. From the results of a literature review, the following procedures were derived, promising for a solution to the problem. For supervised approaches, pre-trained language models may be fine-tuned for a company description matching task. Alternatively, to fine-tuning models that were primarily trained on company descriptions, training on different RL tasks or domains (such as product titles) might provide better results, as indicated by [16]. For completely unsupervised approaches, computing similarity scores based on sentence representations and measures like the cosine distance might provide results comparable to supervised approaches.

Table 2 shows the results of comparing the company descriptions from two example records using Jaro-Winkler, Tf-IDF, and BERT. Jaro-Winkler [14], mostly used for a single word or phrase matching, is not designed for matching natural texts. In our example (table 2) Jaro-Winkler assigns a higher similarity to the non-match tuple than to the match tuple. Document-Vectors such as calculated by Tf-IDF [15] combined with a vector-similarity metric like cosine are not applicable either because the algorithm is word-based and not able to take semantics into account. In our example in table 2, the Soft TF-IDF shows a similarity of 10% and 3% for the two tuples. Embeddings like BERT may solve this issue by incorporating full context information and language understanding while not needing specific training data. Of the three methods, only the BERT based similarity leads to correct classifications in our example (see table 2) because of the semantic information extracted.

Additionally, one could experiment with extracting keywords from the descriptions and match them separately using string similarity metrics that have proven to be working on atomic attributes such as Jaccard or Jaro-Winkler. For example, one could use named entity recognition (NER) models to identify entities described in the texts (like companies, persons, and geographical locations), extract additional descriptive attributes assigned to those entities and try

matching these using string similarity measures. Instead of comparing the entities and attributes extracted themselves, one could also perform classification (either using the full text or extracted attributes as a baseline) to map fixed classes, e.g., sectors. This has strong similarities to traditional matching approaches that primarily use semi-structured data (such as sector or category names that do not follow a predefined taxonomy) and thus should provide good baseline results for matching entities using descriptions. Table 6 shows the preliminary results on the mixed language validation dataset after implementing and evaluating selected approaches mentioned above without further optimization except some tuning on the score thresholds on which a match is predicted. Thus, these models can be regarded as baseline models that more complicated solutions may be compared. For the language models the Huggingface/Transformers library for python [28] has been chosen, because it supports the direct application and further training of different NLP transformer architectures for uniform benchmarking. The experiments that use language models were run on a single NVIDIA GeForce RTX 2080 TI GPU. The training time as well as the validation time for the total dataset have been added to provide a quick overview over the efficiency of the approaches used, since runtime may be a crucial criterion for practitioners and researchers alike. The sklearn [29] accuracy score⁸ has been chosen as a evaluation metrics for measuring the validation accuracy of the different approaches (see table 6). It measures the amount of samples that were correctly predicted (true negatives + true positives) against the total amount of samples. It can be seen, that the (1) fixed choice model that predicts a non-match in every case reaches an accuracy of 55.7% because the validation dataset contains slightly more negative samples due to the extra samples from the manual annotation process present. (2) Jaro-Winkler reaches an accuracy of 55.7% and is therefore not better than a fixed choice model no matter what threshold was chosen. The word based (3) jaccard reached a validation accuracy of 69.3%, although the descriptions in a lot of examples are coming from two different languages (English and German) and thus should not contain the same words in a majority of cases. Experimenting with purely English based pretrained language models (4) BERT and (6) RoBERTa do achieve with 57% and 58% a lower performance than the (3) jaccard approach when the last hidden layer is used in conjunction with the cosine similarity. Using a multilingual (5) BERT the accuracy reaches 78.3%, which was in line with our expectations as we thought a multilingual model should outperform its monolingual counterpart on a multilingual dataset. While the tested (7) XLM specifically was pretrained on English and German texts, it could only outperform the monolingual (4) BERT and (6) RoBERTa by a margin of 2-3%. It could not deliver the same results of the multilingual (5) BERT approach. This shows, that the model architecture clearly has an impact on the performance of this task and further experimentation needs to be done. Using a supervised approach, fine-tuning a (8) BERT model for the specific RL task yielded the best model so far with 80.6% accuracy. Taking the additional amount of work for creating training data and the required model training time into consideration, it is interesting that the accuracy is only marginally better compared to the (5) BERT unsupervised approach that works out of the box. However, the results may improve when further hyperparameter-tuning is applied. Apart from our experiments with single model approaches, combinations of approaches should also be explored in the future. Examples of combining the approaches are shown in

⁸https://scikit-learn.org/stable/modules/generated/sklearn.metrics.accuracy_score.html

Table 6

Preliminary results applying different approaches to the mixed language evaluation set

Approach	Variation	train time	Validation time	Validation Accuracy
(1) Fixed Choice	Always Non-Match	-	< 2 ms.	55.7%
(2) Jaro-Winkler		-	17 sec.	55.7%
(3) Jaccard	Word-based, threshold = 0.05	-	20 ms.	69.3%
(4) BERT unsup.	base English cased, threshold = 0.60	-	10.36 min	57.0%
(5) BERT unsup.	base multilingual cased, threshold = 0.70	-	9.71 min	78.3%
(6) RoBERTa unsup.	base English cased, threshold = 0.97	-	9.63 min	58.0%
(7) XLM unsup.	clm English German, threshold = 0.70	-	5.96 min	60.6%
(8) BERT sup.	base multilingual cased, one epoch	9.5h	10.15 min	80.6%

Figure 2. As mentioned above, using the proposed evaluation set, a supervised model could be trained to determine a similarity score and ultimately classify whether it is a MATCH or not (Approach A). For this approach, multiple experiments might be conducted using different language model architectures (e.g., BERT) and training paradigms (learned from scratch vs. fine-tuning of models trained on other RL tasks like name matching) and preprocessing options. Approach B shows a way to incorporate different models into one matching pipeline by feeding the outputs of one NER and one keyword extraction model into another model that classifies the label (MATCH). This could also be done by language models or by simple string similarity measures like Jaro-Winkler. Combining Approaches A and B yields a pipeline concept in which separate models determine separate matching scores that get averaged to a combined score (Approach C). Additionally, applying unsupervised embedding matching by, e.g., comparing sentence vectors created by a language model by using distance metrics like cosine alongside a supervised model, might also provide interesting insights into which approach combinations work and which do not (Approach D). While it might be the case that the accuracy does not increase by simply using more different models in the pipeline there might be some combinations of approaches that supplement each other to increase information density and therefore accuracy for solving the RL task. That is why various experiments with many combinations should be tested to find the optimal setup. In our overall research environment surrounding these experiments, different data sources with company data are integrated using RL approaches on different data types. The RL system developed does not use the company description, although such texts are prevalent in lots of data sources, as shown in the motivation. We see significant potential in using these descriptions in RL tasks. However, it is unclear what the best setup is (hence the proposal of experiments in section 4) and whether the addition of texts increases the performance of RL systems that already use other attributes to link records. In additional field experiments using the best of the approaches mentioned above, texts will be integrated into the project’s RL system, and its benefit will be measured. A final evaluation of the approaches is to be carried out through A/B testing, which will show whether the new method improves the company integration results. By doing that, we want to measure the actual impact it has to consider descriptions in RL systems to support researchers and practitioners in their decision whether to include those. Testing various combinations with attributes

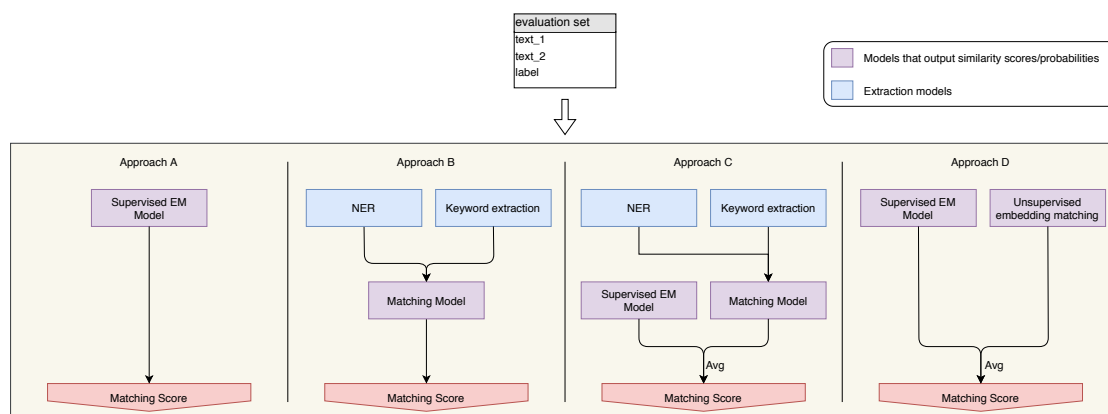


Figure 2: Possible approaches integrating different matching methods into RL pipelines

already used by our RL system (such as names, addresses or legal forms among others), we want to answer not only whether it might be worth using descriptions in the RL process but also in which conditions, e.g. only when neither the attribute name nor address is present.

5. Conclusion and future work

This paper describes the ongoing research work to include natural language text attributes in RL. This goal is to be explored using company descriptions and machine learning methods in the domain of language models. This paper provides a qualitative literature review that shows research gaps and is used to identify potential methods and approaches to include natural language text attributes in RL. Subsequently, a gold standard evaluation and training set has been proposed containing company descriptions from Wikipedia and Crunchbase. Furthermore, a design space exploration was performed to show possible approaches to solve the task of RL using natural language descriptions. Preliminary results have shown, that language models are capable of solving this task and generally perform better than traditional string matching results. The results have shown an interesting opportunity to further optimize the applicability of machine learning methods in RL systems, reducing the amount of human work required in linking data sources to knowledge bases. Within further laboratory experiments, additional proposed methods and approaches and combinations of approaches in RL pipelines will be applied to the gold standard data set and evaluated. Finally, the evaluation will be carried out using the newly developed method to include natural language text attributes within a field study in a real and practice-relevant RL workflow with an industry partner. Our paper contributes to theory and practice by researching natural language text attributes in RL. The improvement of the RL process will optimize data integration in practice. The limitations of our paper also offer opportunities for future research. We focus on the attribute company description while also conceptualizing future solution methods. Further natural language text attributes should be explored and our best approach should be transferred to these attributes and be evaluated.

References

- [1] T. Wrona, P. Reinecke, Wie strategisch sind Algorithmen? Die Rolle von Big Data und Analytics im Rahmen strategischer Entscheidungsprozesse, in: *Logistik im Wandel der Zeit - Von der Produktionssteuerung zu vernetzten Supply Chains*, Springer, 2019, pp. 443–467.
- [2] X. Dong, D. Srivastava, Big data integration, *Synthesis Lectures on Data Management* 7 (2015) 1–198.
- [3] N. Kooli, R. Allesiardo, E. Pigneul, Deep learning based approach for entity resolution in databases, in: *Asian Conference on Intelligent Information and Database Systems*, Springer, 2018, pp. 3–12.
- [4] S. Mudgal, H. Li, T. Rekatsinas, A. Doan, Y. Park, G. Krishnan, R. Deep, E. Arcaute, V. Raghavendra, Deep Learning for Entity Matching: A Design Space Exploration, in: *Proceedings of the 2018 International Conference on Management of Data*, ACM, 2018, pp. 19–34.
- [5] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, *arXiv preprint arXiv:1301.3781* (2013).
- [6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
- [7] K. Karthikeyan, Z. Wang, S. Mayhew, D. Roth, Cross-lingual ability of multilingual bert: An empirical study, in: *International Conference on Learning Representations*, 2019.
- [8] J. Webster, R. T. Watson, Analyzing the past to prepare for the future: Writing a literature review, *MIS quarterly* (2002) xiii–xxiii.
- [9] F. Kruse, A. P. Hassan, J.-P. Awick, J. Marx Gómez, A qualitative literature review on linkage techniques for data integration, in: *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.
- [10] M. Ebraheem, S. Thirumuruganathan, S. Joty, M. Ouzzani, N. Tang, Distributed representations of tuples for entity resolution, *Proceedings of the VLDB Endowment* 11 (2018) 1454–1467.
- [11] L. Li, X. Shang, J. Li, J. Hu, Learning distance metrics for entity resolution, *IEEE Access* 6 (2018) 54900–54909.
- [12] P. Ristoski, P. Petrovski, P. Mika, H. Paulheim, A machine learning approach for product matching and categorization, *Semantic web* 9 (2018) 707–728.
- [13] A. T. Schneider, A. Mukherjee, E. C. Dragut, Leveraging social media signals for record linkage, in: *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 1195–1204.
- [14] A. Y. Sim, A. Borthwick, Record2vec: unsupervised representation learning for structured records, in: *2018 IEEE International Conference on Data Mining (ICDM)*, IEEE, 2018, pp. 1236–1241.
- [15] G. Song, L. Zhang, P. Wang, Entity matching using different level similarity for different attributes, in: *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, IEEE, 2018, pp. 779–782.
- [16] S. Thirumuru., S. A. P. Parambath, M. Ouzzani, N. Tang, S. Joty, Reuse and adaptation for entity resolution through transfer learning, *arXiv preprint arXiv:1809.11084* (2018).

- [17] U. Brunner, K. Stockinger, Entity matching on unstructured data: an active learning approach, in: 2019 6th Swiss Conference on Data Science (SDS), IEEE, 2019, pp. 97–102.
- [18] T. Gschwind, C. Miksovic, J. Minder, K. Mirylenka, P. Scotton, Fast record linkage for company entities, in: 2019 IEEE International Conference on Big Data (Big Data), IEEE, 2019, pp. 623–630.
- [19] D. Javdani, H. Rahmani, M. Allahgholi, F. Karimkhani, Deepblock: A novel blocking approach for entity resolution using deep learning, in: 2019 5th International Conference on Web Research (ICWR), IEEE, 2019, pp. 41–44.
- [20] H. Nie, X. Han, B. He, L. Sun, B. Chen, W. Zhang, S. Wu, H. Kong, Deep sequence-to-sequence entity matching for heterogeneous entity resolution, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019, pp. 629–638.
- [21] A. Primpeli, R. Peeters, C. Bizer, The WDC training dataset and gold standard for large-scale product matching, in: Companion Proceedings of The 2019 World Wide Web Conference, 2019, pp. 381–386.
- [22] C. Zhao, Y. He, Auto-em: End-to-end fuzzy entity-matching using pre-trained deep models and transfer learning, in: The World Wide Web Conference, 2019, pp. 2413–2424.
- [23] Y. Li, J. Li, Y. Suhara, A. Doan, W.-C. Tan, Deep entity matching with pre-trained language models, arXiv preprint arXiv:2004.00584 (2020).
- [24] V. Meduri, L. Popa, P. Sen, M. Sarwat, A comprehensive benchmark framework for active learning methods in entity matching, arXiv (2020) arXiv–2003.
- [25] R. Wu, S. Chaba, S. Sawlani, X. Chu, Zeroer: Entity resolution using zero labeled examples (2020).
- [26] D. Zhang, Y. Nie, S. Wu, Y. Shen, K.-L. Tan, Multi-context attention for entity matching, in: Proceedings of The Web Conference 2020, 2020, pp. 2634–2640.
- [27] P. Mayring, T. Fenzl, Qualitative Inhaltsanalyse, in: Handbuch Methoden der empirischen Sozialforschung, Springer, 2014, pp. 543–556.
- [28] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Huggingface’s transformers: State-of-the-art natural language processing, ArXiv abs/1910.03771 (2019).
- [29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.