CLRL: Feature Engineering for Cross-Language Record Linkage

Öykü Özlem Çakal
Technische Universität Berlin
o.cakal@campus.tu-berlin.de

Mohammad Mahdavi
Technische Universität Berlin
mahdavilahijani@tu-berlin.de

Ziawasch Abedjan
Technische Universität Berlin
abedjan@tu-berlin.de

ABSTRACT

Record linkage aims at identifying duplicate records across datasets. Most existing record linkage techniques have been designed for monolingual datasets. In this paper, we propose a novel approach, CLRL, that links the records in a cross-language setting, where each input dataset is in a different language. CLRL combines monolingual similarity measures with multilingual cross-language word embedding similarities to identify the correspondence of records across datasets. As our experiments show, CLRL outperforms baseline approaches in cross-language data integration settings.

1 INTRODUCTION

Record linkage is one of the most relevant tasks in a data integration process. The goal of record linkage is to identify records from two different datasets that represent the same real-world entity. One of the major challenges in record linkage is to identify similarity heuristics that are effective at approximating the equality of two heterogeneously represented entities [8]. Numerous similarity measures have been proposed so far to capture various similarity levels such as character-based and phonetic-based similarities [3]. These measures are effective in the monolingual setting, where similar words have a high lexical similarity. However, they are often ineffective in a cross-language settings, where each dataset adheres to a different language.

A naive solution to overcome this problem is to first translate one dataset into the other corresponding language and then apply an off-the-shelf record linkage approach. However, this approach suffers from two major problems.

1. *Ambiguity in translation.* Short texts in structured datasets do not usually provide enough context for machine translation models to translate accurately.
2. *Out-of-vocabulary terms.* The machine translation model cannot translate out-of-vocabulary terms, such as the concatenation "firstsight", which is not among the standard language vocabulary.

Motivation example. Table 1 illustrates two movie datasets in English (dataset A) and German (dataset B). Among all the 4 possible pairs of records in $A \times B$, only the second record of the English dataset should be linked to the first record of the German dataset. Traditional record linkage approaches would fail to identify their correspondence as there is no lexical similarity between the title "Forbidden Planet" and the title "Alarm im Weltall". Translating one of the datasets and then applying a traditional record linkage approach also fails because the translation of "Alarm im Weltall" is "Alarm in the Universe"; it is still not lexically similar to "Forbidden Planet".

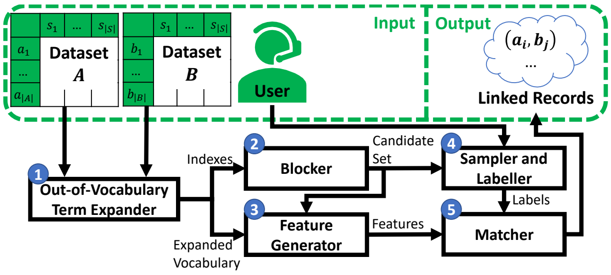


Figure 1: The workflow of CLRL.

Table 1: Two movie datasets in English and German.

ID	Name	Year	ID	Name	Jahr
1	Heat	1995	1	Alarm im Weltall	1956
2	Forbidden Planet	1956	2	Der Pate	1972

However, our approach is able to identify the correspondence by using the latent similarity of these two multilingual short movie titles based on cross-language word embedding models [14]. In particular, our maximum vector similarity feature (see Section 3.2) captures the similarity of the words "Planet" and "Weltall" (means "universe" in German) as these semantically similar words have similar word embedding vectors. □

Contributions. In this paper, we propose a novel approach to link the records across multilingual datasets. To this end, we make the following contributions:

- We design a term expansion scheme (Section 3.1) to expand each out-of-vocabulary term into a set of in-vocabulary terms. In fact, CLRL leverages a three-step policy to expand different types of out-of-vocabulary terms differently.
- We propose an effective set of similarity features (Section 3.2) for cross-language record linkage problem. In addition to state-of-the-art monolingual similarity measures, we include four multilingual similarity measures, adopted from cross-language word embedding models.
- We empirically evaluate our approach on six real-world datasets (Section 4). In particular, we show that CLRL outperforms three existing record linkage approaches in cross-language setting.

2 CLRL OVERVIEW

Figure 1 illustrates the record linkage procedure with CLRL. The multilingual datasets and the user feedback are the input and the set of linked records is the output of the approach.

CLRL adheres to the well-known pipeline of existing record linkage approaches, i.e., preprocessing, blocking, and matching. In addition to standard preprocessing operations, such as value normalization and identifying corresponded attributes, our approach can apply a novel preprocessing step to expand out-of-vocabulary terms into in-vocabulary terms (Step 1). We will detail this step in Section 3.1. In the blocking step, a state-of-the-art blocker is used to generate a set of candidate links between record

© 2019 Copyright held by the owner/author(s). Published in Proceedings of the 22nd International Conference on Extending Database Technology (EDBT), March 26-29, 2019, ISBN 978-3-89318-081-3 on OpenProceedings.org.
Distribution of this paper is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.0.

pairs of the input datasets (Step 2). Then, CLRL generates features for each candidate pair (Step 3). We will detail this step in Section 3.2. Depending on the sampling strategy and user-interaction model, a sample set of candidates are chosen to be labeled by the user as matches or non-matches (Step 4). Finally, a state-of-the-art classifier takes the features and labeled record pairs to classify all record pairs in the candidate set (Step 5).

3 FEATURE ENGINEERING

We first explain how out-of-vocabulary terms are expanded into in-vocabulary terms. Then, we describe our extensive feature set.

3.1 Out-of-Vocabulary Term Expansion

Out-of-vocabulary terms are those terms that could not be found in formal vocabularies and therefore are not translatable. The goal of out-of-vocabulary term expansion is to transform these non-translatable terms into in-vocabulary terms. This way, we can later leverage cross-language word embedding models to capture the similarity of these terms as well. CLRL applies the following steps to the out-of-vocabulary terms.

Morphological checking. CLRL first tries to find morphological variants of the out-of-vocabulary term. In morphological check, the out-of-vocabulary term is transformed into its morphemes (i.e., primitive units), if applicable. For example, the out-of-vocabulary term "firstsight" is transformed into in-vocabulary terms "first" and "sight". We leverage Polyglot Python module [16], which supports 100 languages, to conduct morphological transformations. Note that our morphological checking already covers lighter lemmatization and stemming transformations as well.

Spell checking. Spelling errors could be the emerging cause of many out-of-vocabulary terms. Therefore, in case of failure in morphological transformation, CLRL tries to fix spelling errors. To this end, the approach collects all the in-vocabulary terms whose Damerau-Levenshtein edit distance to the out-of-vocabulary term is less than equal to $\theta_{dist} = 1$. To minimize the risk of replacing an out-of-vocabulary term with the wrong in-vocabulary term, we restrict the threshold to the minimum possible distance, i.e., $\theta_{dist} = 1$, and replace the term only when exactly one in-vocabulary candidate has been found.

Ostrich policy. CLRL ignores transforming all the other out-of-vocabulary terms that could not be transformed by the previous treatments. These out-of-vocabulary terms are mainly numbers (e.g., "2001" in "2001: A Space Odyssey") or named entities (e.g., "Lebowski" in "The Big Lebowski") that do not need any transformation.

3.2 Feature Vector

Each pair of records in the candidate set is mapped to a feature vector that contains all the similarity scores of the two records. Let $A = \{a_1, a_2, \ldots, a_{|A|}\}$ and $B = \{b_1, b_2, \ldots, b_{|B|}\}$ be two relational datasets with different languages, where each $a \in A$ or $b \in B$ is a record. Let $S = \{s_1, s_2, \ldots, s_{|S|}\}$ be the set of mapped attributes in these datasets. Therefore, the data cell $a[s]$ refers to the record $a \in A$ and the attribute $s \in S$. Let f be a similarity function that takes two data cells $a[s]$ and $b[s]$ and returns a similarity score $f(a[s], b[s]) \in [0, 1]$. Therefore, the feature vector of a candidate record pair (a, b) is

$$V(a, b) = [f(a[s], b[s]) \mid \forall f \in F \wedge \forall s \in S], \qquad (1)$$

where F is the set of all the similarity functions. Due to the cross-language setting, the similarity functions should be able to capture, not only the monolingual similarity, but also the multilingual similarity of terms in different languages. That is why, we leverage monolingual and multilingual similarity functions.

3.2.1 *Monolingual Similarity Functions.*

Monolingual similarity functions are typical lexical similarity measures. The goal of incorporating these measures is to capture lexical similarity of named entities, such as "Brad Pitt", which are written similarly in languages with the same scripting system. In particular, we calculate Jaccard, Levenshtein, Jaro, Jaro-Winkler, Needleman-Wunsch, Smith-Waterman, and Monge-Elkan similarity measures [5].

3.2.2 *Multilingual Similarity Functions.*

Multilingual similarity functions capture the similarity of data values across different languages. We leverage cross-language word embedding models [14] to capture the similarity of short multilingual data values of datasets. Word embedding models, such as word2vec [12], learn to map each term into a dense vector in a way that terms with similar context (i.e., surrounding words) have similar vector representations as well. Cross-language word embedding models, such as fastText [7], are a special kind of word embedding models. These models share the same cross-language space for two different languages so that similar words from different languages can have similar vector representations. Thus, a cross-language word embedding model m can take a word w and returns its correspondent vector $m(w)$ in a cross-language shared space. In this shared space, not only monolingual similar words such as "Dog" and "Puppy" would have close vector representations, but also cross-language similar words such as "Dog" and "Hund" (means "dog" in German) would have close vector representations, i.e., $m(\text{"Dog"}) \approx m(\text{"Hund"})$.

Now, let m be the cross-language word embedding model and let $P = \{p_1, p_2, \ldots, p_{|P|}\}$ and $Q = \{q_1, q_2, \ldots, q_{|Q|}\}$ be the sets of words in data cells $a[s]$ and $b[s]$, respectively. We define the following four similarity functions.

Mean vector similarity (MeVS). The Mean of word vectors in a data cell is a representative vector for the whole data cell. Considering each data cell as a set of word vectors, we calculate the cosine similarity of mean vectors of data cells. Formally,

$$MeVS(P, Q) = \text{cosine}\Big(\frac{1}{|P|}\sum_{p \in P} m(p), \frac{1}{|Q|}\sum_{q \in Q} m(q)\Big). \qquad (2)$$

Maximum vector similarity (MaVS). When the data cell contains noisy (i.e., irrelevant) words, it is desirable to represent the data cell by the vector of its most important word. For example, "Forbidden Planet" and "Alarm im Weltall" are the same movie titles in English and German. Since "Weltall" means "universe" in German, $m(\text{"Weltall"})$ is a more accurate representation for the German movie title rather than mean of all the word vectors inside the complete German movie title. This similarity function outputs the maximum cosine similarity between all the pairs of words $P \times Q$. Formally,

$$MaVS(P, Q) = \max_{(p,q) \in P \times Q} \text{cosine}\Big(m(p), m(q)\Big). \qquad (3)$$

Optimal alignment similarity (OAS). If two data cells are matched, they might have an optimal one-to-one alignment of words, where each word in the first data cell corresponds directly to one word in the other data cell. This similarity function looks for such an optimal alignment between words of P and Q, where the sum of similarity scores between aligned word pairs is maximal. Finding the optimal one-to-one alignment of words is the classical assignment problem. We leverage the Hungarian algorithm [9] to find the optimal one-to-one alignment of words in two data cells. Since the aligning needs both P and Q to have the same length of words, the shorter one is padded with arbitrary out-of-vocabulary words, hence inducing dissimilarity. Let us assume that P is the shorter one and its padded version is P'. Formally,

$$OAS(P,Q) = \frac{\sum_{(p,q)\in L(P',Q)} \text{cosine}\Big(m(p), m(q)\Big) \times (|P| + |Q|)}{2 \times |P| \times |Q|}, \tag{4}$$

where $L(P',Q) = \{(p,q) \mid p \in P', q \in Q\}$ is the optimal one-to-one alignment of words in P' and Q. Note that since this similarity function depends on the word count in P and Q, we normalize its value by the harmonic mean of these word counts, as suggested in literature [4].

Maximum alignment similarity (MAS). Two matching data cells might not necessarily have an optimal one-to-one alignment of words. For example, although the English data value "Purchase Price" is matched to the German data value "Kaufpreis", there is no optimal one-to-one alignment for words. Instead, here we have a two-to-one alignment for the words. Thus, in general, we also need a similarity function that can capture the similarity of m-to-n alignments. This similarity function allows each word in the first data value be aligned to the most similar word in the second data value, regardless of whether these two words are already aligned to any other words or not. Formally,

$$\begin{aligned} MAS(P,Q) = \frac{1}{2}\Big(\frac{1}{|P|}\sum_{p\in P} \text{cosine}\Big(m(p), m(q_p^*)\Big) + \\ \frac{1}{|Q|}\sum_{q\in Q} \text{cosine}\Big(m(q), m(p_q^*)\Big)\Big), \end{aligned} \tag{5}$$

where, $q_p^* \in Q$ is the most similar word to word $p \in P$, i.e., $q_p^* = \max_{q\in Q} \text{cosine}\Big(m(q), m(p)\Big)$.

4 EVALUATION

Experimental setup. We evaluate our approach on six real-world datasets, which are described in Table 2. *Universities* and *Universités* contain information of universities around the world in English and French, respectively. An example of matched universities in these datasets is "Technical University of Berlin" and "Université technique de Berlin". *Movies* and *Películas* contain information on movies in English and Spanish, respectively. An example of matched movies in these datasets is "The Godfather" and "El padrino". *Wikipedia Titles* and *Wikipedia-Titel* contain wider domains including titles of Wikipedia pages in English and German, respectively. An example of matched titles in these datasets is "1982 World Snooker Championship" and "Snookerweltmeisterschaft 1982". We extracted these datasets from the DBpedia knowledge base [10]. We leveraged inter-language links inside DBpedia to obtain the ground truth for these datasets, i.e., the pairs of records that are actually linked. We evaluate our approach with precision $P = \frac{\text{the number of correctly identified linked records}}{\text{the number of all outputted linked records}}$, recall $R = \frac{\text{the number of correctly identified linked records}}{\text{the number of all actual linked records}}$, and F_1 measure $F_1 = \frac{2\times P\times R}{P+R}$. We mainly report only the F_1 measure, which combines the precision and recall, due to the space constraints.

Table 2: Datasets.

Name	Language	#Rows	#Common Attributes	Candidate Set Size	#Actual Linked Records
Universities Universités	English French	8758 3957	16	124559	940
Movies Películas	English Spanish	1273 15334	14	59198	72
Wikipedia Titles Wikipedia-Titel	English German	1976 2159	2	51211	83

We apply cross-validation and report the mean and standard deviation of these measures. As the default parameter setting, we setup our approach with all the introduced features. We also leverage fastText [7] as the cross-language word embedding model and XGBoost [1] as the classifier. Our prototype is available online[1].

Effectiveness versus baselines. We compare the effectiveness of CLRL to the following three baseline approaches:

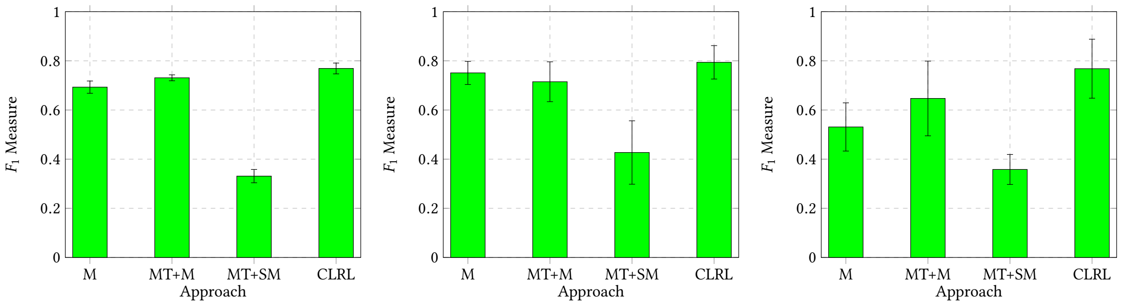
1. **Magellan (M).** Magellan is an end-to-end entity matching system that uses monolingual lexical similarity features to learn links between records [8]. In our experiments, we include the following default set of features: Jaccard, Levenshtein, Jaro, Jaro-Winkler, Needleman-Wunsch, Smith-Waterman, and Monge-Elkan similarity measures.
2. **Machine translation plus Magellan (MT+M).** This approach first translates non-English language datasets into English using the Joshua machine translation toolkit [11] and then applies Magellan on two English datasets.
3. **Machine translation plus semantic matching (MT+SM).** This approach also leverages machine translation to have both datasets in English. However, instead of using Magellan, it applies a monolingual word embedding model to link records [15].

Figure 2 illustrates the effectiveness of CLRL in comparison to these baseline approaches. CLRL always outperforms the other approaches as it leverages a broad set of features to capture monolingual and multilingual similarities. This superiority is more obvious on *Wikipedia Titles/Wikipedia-Titel* datasets, as they contain more linguistically different content. In fact, in university and movie domains there are many named entities such as "Berlin" and "Brad Pitt" that remain the same in many languages. Therefore, even a traditional record linkage approach, such as Magellan, with lexical similarity measures can capture the similarity. However, in Wikipedia title domain there are fewer named entities, hence cross-language techniques are more promising.

Out-of-vocabulary term expansion analysis. Figure 3 illustrates the influence of out-of-vocabulary term expansion on the effectiveness of CLRL. Leveraging out-of-vocabulary term expansion, CLRL has a higher F_1 measure as the similarly functions can capture the similarities with higher recall, i.e., CLRL can identify more linked records. Again, the improvement is higher on *Wikipedia Titles/Wikipedia-Titel* datasets as the out-of-vocabulary terms in these datasets are mainly morphological and spell errors, which are transformed into in-vocabulary terms.

Feature analysis. Table 3 illustrates the effectiveness of CLRL when it leverages different features separately. In general, all the similarity features are informative for the task as CLRL works best with all the features. Furthermore, the proposed multilingual similarity features provide higher F_1 measure than the traditional monolingual features.

[1] https://github.com/BigDaMa/clrl



(a) Universities/Universités (b) Movies/Películas (c) Wikipedia Titles/Wikipedia-Titel

Figure 2: Effectiveness of different approaches on different datasets.

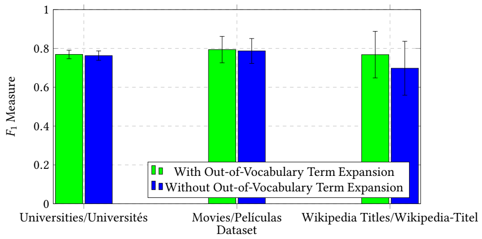


Figure 3: F_1 measure of CLRL with and without out-of-vocabulary term expansion.

Table 3: F_1 measure of CLRL with different feature groups.

Feature Name	Universities/Universités	Movies/Películas	Wikipedia Titles/ Wikipedia-Titel
Monolingual	0.69 ± 0.02	0.75 ± 0.04	0.53 ± 0.10
MeVS	0.70 ± 0.01	0.76 ± 0.08	0.67 ± 0.09
MaVS	0.71 ± 0.02	0.79 ± 0.09	0.60 ± 0.13
OAS	0.72 ± 0.02	0.75 ± 0.03	0.59 ± 0.10
MAS	0.75 ± 0.01	0.80 ± 0.04	0.70 ± 0.14
Full	**0.77 ± 0.02**	**0.80 ± 0.06**	**0.77 ± 0.12**

5 RELATED WORK

Cross-language record linkage. There are only few pieces of work on cross-language record linkage because this is relatively a new topic. Song et al. [15] translated the Japanese datasets into English and then applied monolingual word embedding models to identify linked records. As shown in the experiments, CLRL outperforms this approach because of two reasons. First, CLRL does not rely on direct translation of datasets, which can be ambiguous as explained earlier. Second, instead of only one monolingual word embedding-based similarity feature, CLRL leverages various monolingual and multilingual similarity features to capture the similarities of multilingual records more accurately.

Record linkage. Numerous works have tackled the similarity representation challenge of record linkage task by different similarity measures [3]. In addition to these common monolingual similarity measures, CLRL leverages multilingual word embedding-based similarity measures as well-suited similarity features for cross-language setting. We showed the benefit of the new similarity features in our experimental comparison.

Cross-language matching. Cross-language matching has been mainly studied for unstructured text data in tasks such as information retrieval [6] and entity matching [13]. While the entity is usually surrounded with a rich context of words in these tasks, in structured datasets the texts are mainly short, which make the cross-language matching task more challenging. That is why, CLRL leverages cross-language word embedding models to capture the semantic similarity of short multilingual texts accurately.

Out-of-vocabulary term expansion. Out-of-vocabulary term expansion has been addressed for the record linkage problem using the top-K co-occurring words with the out-of-vocabulary term [2]. CLRL does not hold any assumption on the term frequency of the out-of-vocabulary terms in the dataset.

6 CONCLUSION

We addressed the problem of cross-language record linkage. In addition to the monolingual similarity measures, we leverage four novel cross-language word embedding-based similarity measures. As our experiments show, CLRL outperforms three record linkage baseline approaches in cross-language setting. In future, we plan to extend the blocking step. Since the records are in different languages, simple blocking heuristics, such as having a word in common, do not work effectively in cross-language setting.

ACKNOWLEDGMENTS

This work has been partially funded by the German Ministry for Education and Research as BBDC II (01IS18025A).

REFERENCES

[1] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *SIGKDD*. 785–794.

[2] Muhammad Ebraheem et al. 2017. DeepER–deep entity resolution. *arXiv preprint arXiv:1710.00597*.

[3] Ahmed K Elmagarmid et al. 2007. Duplicate record detection: A survey. *TKDE* 19, 1, 1–16.

[4] Goran Glavaš et al. 2018. A resource-light method for cross-lingual semantic textual similarity. *Knowledge-Based Systems* 143, 1–9.

[5] Wael H Gomaa and Aly A Fahmy. 2013. A survey of text similarity approaches. *IJCA* 68, 13, 13–18.

[6] Gregory Grefenstette. 2012. *Cross-language information retrieval.* Springer Science & Business Media.

[7] Armand Joulin et al. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *EMNLP*. 2979–2984.

[8] Pradap Konda et al. 2016. Magellan: Toward building entity matching management systems. *PVLDB* 9, 12, 1197–1208.

[9] Harold W Kuhn. 1955. The Hungarian method for the assignment problem. *Naval research logistics quarterly* 2, 1-2, 83–97.

[10] Jens Lehmann et al. 2015. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* 6, 2, 167–195.

[11] Zhifei Li et al. 2009. Joshua: An open source toolkit for parsing-based machine translation. In *StatMT*. 135–139.

[12] Tomas Mikolov et al. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*. 3111–3119.

[13] Daniel Rinser et al. 2013. Cross-lingual entity matching and infobox alignment in Wikipedia. *IS* 38, 6, 887–907.

[14] Sebastian Ruder et al. 2017. A survey of cross-lingual word embedding models. *arXiv preprint arXiv:1706.04902*.

[15] Yuting Song et al. 2016. Cross-language record linkage using word embedding driven metadata similarity measurement.

[16] Sami Virpioja et al. 2013. Morfessor 2.0: Python implementation and extensions for Morfessor Baseline. Aalto University.