LOREM

Language-consistent Open Relation Extraction from Unstructured Text

Harting, Tom; Mesbah, Sepideh; Lofi, Christoph

DOI
10.1145/3366423.3380252

Publication date
2020

Document Version
Final published version

Published in
Proceedings of the The Web Conference (WWW)

LOREM: Language-consistent Open Relation Extraction from Unstructured Text

Tom Harting
Sepideh Mesbah
Christoph Lofi
tomharting@gmail.com
s.mesbah@tudelft.nl
c.lofi@tudelft.nl
Delft University of Technology
Delft, The Netherlands

ABSTRACT

We introduce a Language-consistent multi-lingual Open Relation Extraction Model (LOREM) for finding relation tuples of any type between entities in unstructured texts. LOREM does not rely on language-specific knowledge or external NLP tools such as translators or PoS-taggers, and exploits information and structures that are consistent over different languages. This allows our model to be easily extended with only limited training efforts to new languages, but also provides a boost to performance for a given single language. An extensive evaluation performed on 5 languages shows that LOREM outperforms state-of-the-art mono-lingual and cross-lingual open relation extractors. Moreover, experiments on languages with no or only little training data indicate that LOREM generalizes to other languages than the languages that it is trained on.

CCS CONCEPTS

• **Computing methodologies** → **Information extraction**; Machine learning approaches; • **Information systems** → *Data extraction and integration.*

KEYWORDS

open domain relation extraction, multi-lingual relation extraction, text mining

ACM Reference Format:
Tom Harting, Sepideh Mesbah, and Christoph Lofi. 2020. LOREM: Language-consistent Open Relation Extraction from Unstructured Text. In *Proceedings of The Web Conference 2020 (WWW '20), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3366423.3380252

1 INTRODUCTION

Extracting relationships between entities from text is a core building block for (semi-)automatically creating structured knowledge bases. Relation extractors focusing on lexical features and smaller sets of relationship types have shown to be effective, especially in defined domains like bio-medical [9, 18] or law. However, they struggle in less focused applications like general-purpose Web or Social Media mining which are not restricted in relation type or language used. In this paper, we target this use case with a novel open relation extraction model which is also coping with multi-linguality.

Open Relation Extraction (ORE) is defined as the process of discovering arbitrary semantic connections between entities in unstructured texts [5]. Given an input sentence such as *"Turing was born in England in 1912"* and two entities like <*Turing, England*>, an ORE system should extract a sub-string which entails the semantic relation between the two entities (i.e. *"was born in"*).

Initially, ORE research focused on training sequence tagging models by utilizing external NLP tools (such as POS taggers) and manually defined lexical and syntactic features [1, 6, 7, 16]. The dependency on external NLP tools results in error propagation. Also, most of these tools are developed for English only hindering the adoption of ORE algorithms to other languages. Although being a rough estimate, various cross-over studies imply that around 70% of the internet is written in languages other than English [20]. This indicates a need for more generic, language-agnostic ORE models. Recent approaches [4, 12, 19, 23] employed deep neural networks to automatically learn relation patterns from large training sets to tackle the problem of manually defining features and language structures for multiple languages. However, they still require additional NLP tools for pre-processing text such as translators or dependency parsers, thus limiting easy extension to new languages.

Our goal is to exploit similarities and pattern consistencies which exist between many natural languages to replace those language-specific external tools. Recently, Relaxed Cross-domain Similarity Local Scaling (RCSLS) [13] was presented, a word embedding alignment approach which exploits the inter-dependencies between any two languages and maps all monolingual embeddings into a shared multilingual embedding space. In a similar fashion, we leverage existing pre-trained multilingual word embeddings (which are currently available for 44 languages[1]). The intuition behind these efforts is that some languages share common ancestry, and thus exhibit similarities in grammar and vocabulary. We therefore assume that also their trained relation extractors can support each other, which is especially valuable for use cases where a well-trained model is available, but relation extraction is required for a resource-scarce language. For example, we can show that a richly trained

WWW '20, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7023-3/20/04.
https://doi.org/10.1145/3366423.3380252

[1] https://fasttext.cc/

English relation extraction model (for which many manually annotated training corpora are available) can significantly boost the performance of a poorly trained Dutch model (for which only very few training samples are available.)

Based on this intuition, we present LOREM, a model that harvests information that is consistent over languages for Open Relation Extraction. LOREM depends only on monolingual ORE training data and multilingual word embeddings, it can thus be easily extended to new languages.

We make the following contributions:

- We introduce a Language-consistent Open Relation Extraction Model (LOREM). To the best of our knowledge, LOREM is the first open relation extractor that utilizes language-consistent relation structures to improve open relation extraction performance across multiple languages. In addition LOREM does not depend on language-specific knowledge or external NLP tools such as translators or dependency parsers, thus allowing for easy expansion to new languages.
- To the best of our knowledge, we are the first to employ multilingual, aligned word embeddings as the input of a multilingual relation extractor. Our experiments show that this improves the performance over using conventional monolingual word embeddings.
- We present experimental results on five high-resource languages showing that LOREM outperforms state-of-the-art mono-lingual and cross-lingual open relation extractors. Additionally we present experiments on no- and low-resource languages which demonstrate the ease and effectiveness of expanding LOREM to additional languages. This shows that language consistency can not only boost extraction performance for low-resource languages (like Dutch which can benefit from English), but also high-resource languages (like a well-trained English model which still benefits slightly from e.g. a French one),

The source code of LOREM is made publicly available on https://github.com/tomharting/LOREM.

2 RELATED WORK

From the literature, we identify two paradigms; *closed* and *open* relation extraction. For the closed paradigm, the goal is to classify a sentence with respect to a pre-defined set of relation classes. Banko et al. [2] argue that requiring pre-defined relation classes is too limiting for many real-world applications. To alleviate this limitation, they propose the open relation extraction (ORE) paradigm. The vast majority of ORE research is presented for the English language. Although multilingual methods were proposed, they either depend on bilingual training data or solely work in the closed relation extraction domain.

2.1 English Open Relation Extraction

EORE was first introduced by Banko et al. [2]. Conventional models use lexical and syntactic features that rely on external NLP tools and language-specific relation structures. To avoid error propagation by these external tools and alleviate the burden of designing manual features, multiple neural open relation extractors were proposed [4, 12, 19]. Jia et al. [12] present one of the current state-of-the-art model called NST (Neural Sequence Tagger). They define a tagging scheme and predict a tag for each word in the input sentence. For this purpose, they jointly train a CNN and bi-LSTM. The output of these models is fed into a final CRF layer to end up with the final prediction. Their experiments show that CNNs and LSTMs provide complementary information for the RE task.

Even though recent research efforts yield state-of-the-art results for the ORE task by utilizing neural network based models, these works are solely focused on the English language and will encounter two weaknesses when applied in a multilingual setting. First, the vast majority of these systems use external NLP tools such as PoS-taggers and dependency parsers [1, 6, 7, 16] and need to be adapted to use tools for the given language, which is a non-trivial process. Second, EORE would fail to exploit information that is present over multiple languages (language-consistent patterns). Both of these weaknesses are addressed by two different multilingual RE techniques; cross-lingual RE and language-consistent RE. Cross-lingual systems try to extract relations from a source language by exploiting information and systems from a target language, thereby removing the need for a labelled training set or NLP tools in the source language. On the other hand, language-consistent systems exploit information that is present in multiple languages.

2.2 Cross-lingual Open Relation Extraction

Cross-lingual approaches can be used when we need to extract relations from a source language for which we do not have a labelled training set. We do however need to possess either a performant translator [8] or a sufficiently large bi-text corpus between English and the source language [24]. In recent years multiple cross-lingual approaches are introduced [8, 23, 24]. Typically, a cross-lingual system translates the source language into the target language (e.g. English) and employs an existing relation extractor. In an effort to relax the translator assumption and to tailor the translator to the RE task at hand, Zhang et al. [23] present their joint Machine Translation/Information Extraction (MT/IE) system. Instead of first translating the source text and then applying a relation extractor, they jointly train a machine translation and relation extraction model. The translator assumption is replaced by the assumption that a bi-text corpus (e.g. a corpus of Chinese sentences and their aligned English translations) is available.

2.3 Language-consistent Relation Extraction

To remove the dependency on bi-text corpora or translators and to introduce a mechanism for exploiting information that is consistent over languages, language-consistent relation extraction was proposed. Language-consistent RE literature [14] assumes that relation patterns in sentences are substantially consistent between different languages. This assumption can be exploited to train a single model which gathers information from multiple languages. In the previous section, we have seen that cross-lingual relation extractors exist for the open RE domain. In contrast, language-consistent relation extractors are currently solely proposed for the closed domain [14, 21].

Wang et al. [21] train a separate language-individual model for every language and one language-consistent model on all languages in the closed RE paradigm. By combining both models, they can

utilize relation patterns that are specific to languages as well as patterns that are consistent over languages. To ensure that the representations of sentences are aligned over multiple languages, they use an adverserial training approach. To obtain the same latent consistency among languages in similar NLP tasks, multilingual word embeddings were proposed [3, 13] which are trained with the specific goal to align similar words over multiple languages. Multilingual word embeddings use information from multiple high resource languages to create a shared embedding space in which also low-resource language can be represented.

Our work is inspired by Jia et al. [12] and Wang et al. [21]. In contrast, LOREM utilizes language-consistent information within the Open Relation Extraction domain and employs multilingual embeddings [13] for multilingual relation extraction. Our approach does not rely on any external NLP tools or additional bilingual training data, ensuring low-cost extendibility to new languages.

3 LOREM: LANGUAGE-CONSISTENT OPEN RELATION EXTRACTION MODEL

In a nutshell, the idea behind our Language-consistent Open Relation Extraction Model is to start with several language-individual models for each required language. This means that for each language, at least some training data needs to be available (however, as we can show in our experiments for the Dutch language, it can be sufficient to have only a few hundred training samples available which one of the authors could easily provide by himself.) In Figure 1, one language-individual model (to the left) is depicted, but to take full advantage of LOREM, several of such models should be available. We base these models on Neural Sequence Tagging (NST) [12], a recent state-of-the-art approach for (mono-lingual) open relation extraction.

To exploit consistencies between the languages available to the system, we additionally train a language-consistent model using all available languages. The techniques for combining the individual models and the language-consistent model is inspired by AMNRE [21] (a model for language-consistent relation extraction, which is strictly limited to a closed domain of few relationship types. However, we changed the workflow of AMNRE considerably to work with the NST models, e.g. by switching to multilingual embeddings). In the current version of LOREM, only one language-consistent model for any number of languages is used. But as discussed in the conclusions, we see potential for having several of different language-independent models which are trained on selected subsets of the available languages (e.g., those languages which share structural similarities / ancestry).

3.1 Input Embeddings

An input sentence is encoded using two different types of pre-trained word embeddings, one for use with the language-consistent model and one for use with the language-individual models. For the language-individual models, we use conventional pre-trained word embeddings. In Figure 1, these embeddings are represented in blue on the left. The training sentences of the language-individual model all come from the same language, and we expect that this model finds relation structures that are specific to that individual language.

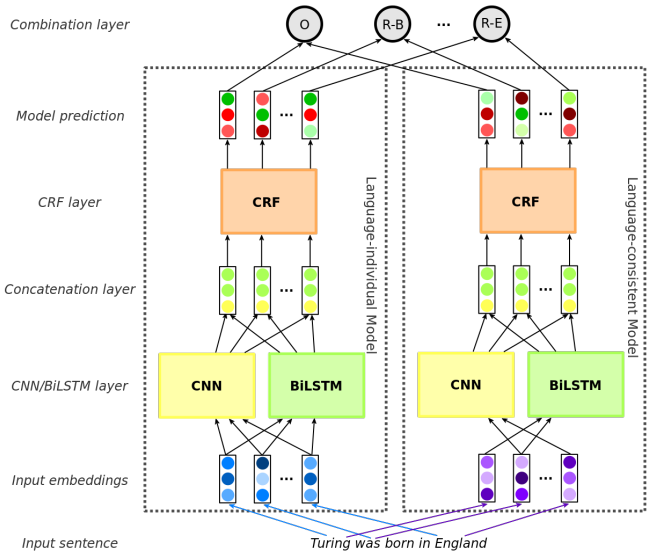


Figure 1: Architecture of our Language-consistent Open Relation Extraction Model (LOREM).

In order to achieve latent consistency among languages, we pioneer the use of multilingual embeddings for the language-consistent model [13]. By using embeddings that are aligned over languages, we hypothesize that we can ease the burden of the CNN/BiLSTM layer to extract language-consistent patterns. Here, the intuition is that the multi-lingual embedding prevents language-specific clusters in the embedding space (such clusters naturally happen when using multiple mono-lingual embeddings). Thus, related or similar words should be close no matter their original language which supports discovery of language-consistent patterns. Note that we use pre-trained embeddings in our current version of LOREM. In scenarios where such dependencies are undesired, such embeddings could also be custom-learned during system setup.

In Figure 1, these embeddings are represented in purple on the right. For this model, the training sentences come from multiple languages. Thus, we expect this model to extract relation patterns consistent over these language.

In addition to word embeddings, entity tag vectors are added to the input. These are simple one-hot encoded vectors which indicate if the current word is part of the first, second or no relation entity. Please note that in contrast to the NST model, we do not use Part-of-Speech tags since these introduce a dependency on PoS-taggers.

The input sentence is represented as a k-dimensional embedding sequence $\mathbf{x} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_n\}$, where $\mathbf{w}_t$ is the representation of the t^{th} word of an input sentence that has n words. Here, $k = k_i + k_c$, k_i and k_c are the dimensonalities of the language-individual and -consistent model input respectively. $k_i = k_{mono} + k_e$ and $k_c = k_{multi} + k_e$, where k_{mono} is the dimensionality of the monolingual word embedding, k_{multi} of the multilingual word embedding and k_e of the entity tag vector.

3.2 NST Layers

The next four layers (CNN/BiLSTM, concatenation, CRF, model prediction) are identical to the NST model. We shortly reiterate the NST model's general architecture, a more detailed description can be found in the original NST paper [12]. Relational words tend to occur in the neighbourhoods of entities. Therefore, certain parts of the input sentence might have a higher chance of containing relation words than others. A CNN is used to capture this local feature information from the input sentence. At the same time, a bidirectional LSTM is used to capture the forward and backward context of each word, including long-distance relations. By concatenating the outputs of the CNN and the forward and backward pass of the LSTM, a continuous representation of each word in the input sentence is formed. Next, these representations are used as the input for a straightforward CRF layer, which tags a word using the NST tagging scheme.

Tag	Meaning
R-S	Single word relation sub-string.
R-B	Beginning of relation sub-string.
R-I	Inside the relation sub-string.
R-E	Ending of relation sub-string.
O	Outside the relation sub-string.

Table 1: NST tagging as proposed by Jia et al. [12].

The NST tagging scheme consists of five possible relation tags, which can be found in Table 1. The sentence *"Alan Turing was born in England."* should be tagged as follows; *"$Alan_O$ $Turing_O$ was_{R-B} $born_{R-I}$ in_{R-E} $England_O$ $._O$"*.

The output of the NST layers are two prediction sequences $\mathbf{y}_{ind} = \{\mathbf{i}_1, \mathbf{i}_2, ..., \mathbf{i}_n\}$ and $\mathbf{y}_{con} = \{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_n\}$, where $\mathbf{y}_{ind}$ contains the predictions of the language-individual model and $\mathbf{y}_{con}$ contains the predictions of the language-consistent model. $\mathbf{i}_t$ and $\mathbf{c}_t$ are the 5-dimensional prediction vectors of the language-individual and -consistent models respectively. For the original NST model, these are binary vectors which contain a 1 for the predicted tag and a 0 for all other tags. After our alteration, these vectors contain a probability score for each of the possible relation tags. This allows us to fittingly combine the predictions of the language-individual and -consistent models in the next layer.

3.3 Combination Layer

In the last layer, we define the final probability sequence by $\mathbf{y} = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_n\}$ with

$$\mathbf{p}_t = \mathbf{i}_t \odot \mathbf{c}_t, \tag{1}$$

for the t^{th} word in the input sentence[2]. The output tag sequence is defined by $\mathbf{z} = \{z_1, z_2, ..., z_n\}$ where

$$z_t = \arg\max_j \mathbf{p}_{tj} \tag{2}$$

and where $\mathbf{p}_{tj}$ is the j^{th} element of $\mathbf{p}_t$.

LOREM might (rarely) yield tag sequences which are invalid. This is a common issue with sequence taggers, including also vanilla NST. For example, the tag for a single word relation (*R-S*) can not be followed by a tag for the end of a multi-word relation (*R-E*). In this case, the first tag could be changed to *R-B* to form a valid tag sequence. We create two different versions of LOREM, LOREM_{clean} which alters invalid sequences to valid sequences and LOREM which allows invalid sequences. To create LOREM_{clean}, we transfer the predicted tags to binary tags (*R* if the word is in the relation, *O* if it is not). Next, we specify the *R* tags so that the first *R* occurrence in a sentence will become *R-B* for a multi-word relation and *R-S* for a single-word relation. Similarly, the last *R* occurrence will become *R-E* and the middle *R* occurrences will become *R-I* for a multi-word relation. Please note that this approach solely influences the specific relation tag that is given to a word, it does not influence whether a word is tagged as being part of the relation or not.

[2] $\odot$ is used as the Hadamard product.

4 EXPERIMENTS

We present experimental results investigating the behaviour of LOREM and its sub-models guided by the following hypotheses:

- *H1:* For *high*-resource (i.e. 100k+ sentences with tagged open relations) languages, LOREM outperforms state-of-the-art monolingual open relation extractors (including NST) by additionally harvesting language-consistent relation patterns from multilingual texts.
- *H2:* Multilingual word embeddings improve the performance of the language-consistent sub-model, and thereby the performance of LOREM by introducing a latent consistency among languages.
- *H3:* For *low*-resource (in our case ~750 tagged sentences) and *no*-resource (i.e. no sentences with tagged open relations) languages, our approach is able to outperform language-individual models by harvesting language-consistent relation patterns from multilingual texts and by utilizing models of languages that have a similar origin.

Our model uses the hyper-parameters that were proposed by Jia et al. [12] for their NST model. We evaluate the performance of our approach using precision, recall and F_1-score.

4.1 Datasets

Information about the used training and test data is presented in Table 2. We used data from the following datasets, covering English, Spanish, French, Hindi, Russian, Italian, and Dutch.[3]

WMORC [8] WMORC contains manually annotated open relation extraction data for 3 languages (WMORC_{human}) and automatically tagged (and thus less reliable) relation data for 61 languages, created using a cross-lingual projection approach (WMORC_{auto}). The sentences are gathered from Wikipedia.

NeuralOIE [4] English dataset created by using only high-confidence extractions of an existing relation extractor [15] from Wikipedia sentences.

ClausIE [6] Three manually annotated English test sets from Wikipedia and New York Times sentences. In line with existing literature, we present averaged results over these three test sets for the English language.

[3] We selected these languages, since these are the only languages for which we could find openly available test data.

			High			**No**	**Low**
	English	Spanish	French	Hindi	Russian	Italian	Dutch
# Training sentences	576,462	429,413	468,625	280,815	550,720	0	750
# Test sentences	2,191	246	512	622	573	10,000	100
Origin training data	NeuralOIE	WMORC$_{auto}$	WMORC$_{auto}$	WMORC$_{auto}$	WMORC$_{auto}$	-	WMORC$_{auto}$
Origin test data	ClausIE	RWP	WMORC$_{human}$	WMORC$_{human}$	WMORC$_{human}$	WMORC$_{auto}$	MC

Table 2: Description of the datasets used in our experiments for high-, no- and low-resource languages. Legend: *RWP* – Raw Web/Parallel En-Sp; *MC* – Manually Created

Raw Web/Parallel En-Sp [25, 26] Two manually annotated Spanish test sets from school text book and web page sentences.

Custom For Dutch, we created our own test set by having a native speaker tag 100 random Dutch Wikipedia sentences (since the Dutch sentences contained in WMORC$_{auto}$ seemed to be of too low quality to be used for testing due to their automatically generated nature).

The size of our high-resource training sets (En, Sp, Fr, Hi, Ru) is comparable to the dataset used in the original NST paper [12]. Moreover, early tests did not show substantial benefits of adding more data after this point. We approach Dutch from a low-resource scenario, so we only sample 750 Dutch sentences from WMORC$_{auto}$ for training. We don't use any Italian training data, since Italian is used as a no-resource language in our experiments (i.e. for Italian, there is no language-individual NST model available during the evaluations, only the language-consistent one). For training the language-consistent model, we sample the high-resource datasets presented in Table 2, so that the combined set of all five languages contains 450,000 - 550,000 training sentences. The selected samples are balanced across these languages. This way, we can make a fair comparison between the language-individual and -consistent sub-models since they are trained on the same amount of training data.

Given the very limited scope of existing multilingual open relation extraction literature, there are only very few results presented for these datasets ('Origin test data' in 2) . Moreover, these were the only publicly available ORE test sets we could find for non-English languages. The Italian test set is created by sampling 10,000 sentences from WMORC$_{auto}$. Since these sentences are automatically tagged, we do expect a higher noise level than in the manually tagged test sets.

For the language-individual model, we use FastText word embeddings [11] which are trained on Common Crawl and Wikipedia dataset. For the language-consistent model, we use pre-trained multilingual embeddings which are released by FastText [13] for 44 languages. The vectors were trained on a Wikipedia dataset. The dimensionality of both embeddings is 300.

4.2 Comparison Methods

During our experiments, we compare LOREM to a range of previously proposed methods. For English, we compare LOREM to the same baseline systems that were used during the evaluation of the NST model by Jia et al. [12]. These include:

NST$_{no-PoS}$ [12] The NST model forms the underlying model of LOREM, yet there are differences between the two. The original NST model does not contain a language-consistent part. We present the results for NST without PoS-tags for a fair comparison.

Reverb [7] Reverb exploits syntactic and lexical constraints on binary relations expressed by verbs.

OLLIE [16] This model designs complex patterns using syntactic processing (e.g. dependency parsers).

ClausIE [6] ClausIE exploits linguistic knowledge about English grammar to detect and identify clauses and their grammatical function.

Open IE-4.x [15] This is a combination of a rule-based Open IE system and a system which analyzes the hierarchical structure between semantic frames to construct multi-verb open relation phrases.

For Spanish, we compare LOREM to;

ExtrHech [26] A system based on syntactic constraints over PoS-tag sequences targeted at Spanish.

ArgOE [10] ArgOE uses dependency parsers to extract a set of propositions for different argument structures.

Finally, we compare LOREM to a **cross-lingual system** presented by Faruqui et al. [8] which utilizes a translator and an English ORE system.

5 EXPERIMENTAL RESULTS

H1: LOREM for High-resource Languages

Table 3 contains the experimental results of LOREM and the comparison methods on five different high-resource test languages. We find that both LOREM models outperform all English baseline systems in terms of recall and F_1-scores. Focusing on the comparison with the NST model, we find that LOREM outperforms NST on precision, recall and therefore F_1-score. The high F_1-scores of our LOREM models are mainly due to the higher recall scores, compared to other systems. LOREM achieves the best presented F_1-score on the ClausIE datasets when PoS-tags are not used. We also find that LOREM achieves comparable results to the NST baseline that does include PoS-tags which obtains 0.869, 0.735 and 0.796 in terms of precision, recall and F_1-score [12]. However, LOREM does have the advantage that it does not rely on external NLP tools and can therefore be more easily extended to new languages.

Next, we compare LOREM to two Spanish open relation extractors. It is important to note that both existing models heavily rely on semantic constraints and external NLP tools. For ArgOE the authors only present a precision score. The results show that LOREM is

	English			Spanish			French			Hindi			Russian		
Model	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Our work															
LOREM	.801	.757	**.782**	.615	.522	.564	.783	.715	**.747**	**.900**	.598	**.719**	**.762**	.719	.740
LOREM$_{clean}$	.782	**.765**	.774	.585	.547	.564	.726	**.729**	.727	.687	**.618**	.651	.709	.726	.718
Language-ind.	.796	.747	.771	.595	.498	.541	.781	.693	.735	.878	.540	.667	.755	**.741**	**.748**
Language-con.	.792	.734	.762	.583	.471	.521	.733	.673	.702	.813	.566	.667	.712	.690	.701
English															
NST$_{no-PoS}$	.783	.708	.744	-	-	-	-	-	-	-	-	-	-	-	-
Reverb	.641	.162	.259	-	-	-	-	-	-	-	-	-	-	-	-
OLLIE	**.985**	.242	.389	-	-	-	-	-	-	-	-	-	-	-	-
ClausIE	.801	.531	.638	-	-	-	-	-	-	-	-	-	-	-	-
Open IE-4.x	.792	.331	.467	-	-	-	-	-	-	-	-	-	-	-	-
Spanish															
ExtrHech	-	-	-	**0.710**	**0.595**	**0.647**	-	-	-	-	-	-	-	-	-
ArgOE	-	-	-	0.500	-	-	-	-	-	-	-	-	-	-	-
Cross-lingual															
Faruqui et al.	-	-	-	-	-	-	**0.816**	-	-	0.649	-	-	0.635	-	-

Table 3: Results of LOREM, its sub-models and existing models. Bolds indicate the best values per language.

outperformed by ExtrHech on the Spanish datasets. It does however achieve a higher precision than ArgOE. Even though the evaluation results are not quite as high as the current state-of-the-art model, LOREM does have the big advantage that a user does not have to manually define semantic constrains. The drop in performance is a clear trade-off with the time and labor involved in building the ExtrHech model.

We now turn our attention towards the three remaining test languages. To the best of our knowledge, there exists only one system for which results are published on the WMORC$_{human}$ test set, being the cross-lingual model by Faruqui et al. [8]. For this model, the source code is not available and only precision scores are presented. We find that the cross-lingual model slightly outperforms LOREM in terms of the French precision score. However, LOREM clearly outperforms the cross-lingual model on both Hindi and Russian. This might be caused by the fact that the cross-lingual model is heavily dependent on a translator from English to the target language and an existing English relation extractor. LOREM eliminates this dependency by introducing a language-consistent component. The results indicate that this improves the generalizing capabilities over languages, providing prove for the validity of hypothesis 1.

In order to investigate how well each submodel in LOREM performs, we presented the results obtained by the sub-models in Table 3. LOREM generally outperforms both the language-individual and -consistent model, showing the merit of combining these sub-models. This falls in line with the conclusions presented by Wang et al. [21] for the closed domain.

In addition to these findings, we also observe a returning pattern between LOREM and LOREM$_{clean}$. For all languages, LOREM achieves higher precision and F_1-scores, indicating a better overall performance. However, cleaning the prediction results does consistently improve the recall of the model. Thus, we conclude that LOREM generally outperforms LOREM$_{clean}$, yet LOREM$_{clean}$ should be used when recall is crucial for the application domain.

Another, somewhat surprising, observation from Table 3 is the reasonably good performance of the language-consistent model, given the fact that this sub-model is not trained on one specific language. From these results, we wondered if relation structures truly differ a lot between languages. It could be the case that a language-individual model already performs reasonably well on other languages, eliminating the need for a language-consistent model. To test this hypothesis, we compare the average results of the language-consistent model over all five languages to the average results of the language-individual models on these languages. The results are presented in Table 4. The results clearly counteract the hypothesis, showing the merit of a language-consistent model over simply using one language-individual model for every language.

Model	**P**	**R**	**F_1**
Language-consistent	**.727**	**.627**	**.671**
English language-individual	.393	.317	.347
Spanish language-individual	.586	.390	.455
French language-individual	.679	.464	.543
Hindi language-individual	.266	.110	.138
Russian language-individual	.632	.483	.546

Table 4: The average prediction results of the language-consistent model and language-individual models on all test languages. The bolds indicate the best values.

H2: Multi VS Monolingual Embedding

Current multilingual relation extraction literature utilizes monolingual word embeddings to encode sentences of different languages. However, we expect the model to extract patterns that are consistent over languages. Therefore, the model should ignore the language in which an input word is written. Naturally, aligning

these word embeddings of languages would ease the burden of the language-consistent model to extract language-consistent relation patterns.

To examine this hypothesis, we compare the results obtained by using both non-aligned (monolingual) and aligned (multilingual) word embeddings. All other variables, such as training sets, model architectures and parameters, remain the same. In Figure 2, we present the results of this experiment. Additionally, we provide the impact of both approaches on the full LOREM model, showing that improvements for the language-consistent sub-model indeed lead to improvements of the full model. We observe that the aligned word embeddings yield better performance on every language for both the language-consistent sub-model and the full LOREM model in terms of F_1-score. Given these test results, we can confirm the validity of hypothesis 2.

H3: LOREM for Low/No-resource Languages

The evaluation results for low- and no-resource languages are shown in Table 5 and 6. If no open relation extraction training data is available for a certain language, our model can still be utilized in three possible ways: 1) we can use the language-consistent sub-model trained on other languages, 2) we can use a language-individual model of a language that has a similar origin to the current language 3) or we can combine both into a full LOREM model. If we also have a very small training set of around 750 sentences (for the low-resource scenario), we can additionally train a language-individual model using it.

	Dutch			Italian		
Model	P	R	F_1	P	R	F_1
Language-con.	.705	**.633**	.667	.506	**.342**	**.408**
English	.655	.582	.616	.293	.232	.259
Spanish	.441	.306	.361	.435	.203	.277
French	.685	.510	.585	.352	.217	.268
Hindi	.000	.000	.000	.362	.029	.054
Russian	.703	.265	.385	.393	.164	.232
LOREM	**.744**	.622	**.678**	**.554**	.246	.341
LOREM$_{clean}$	.663	.622	.642	.383	.287	.328

Table 5: (no-resource) Results of the language-consistent model, language-individual models and LOREM on the Dutch and Italian test sets.

Model	**P**	**R**	**F_1**
Language-individual	**.786**	.444	.568
LOREM	.753	**.646**	**.696**

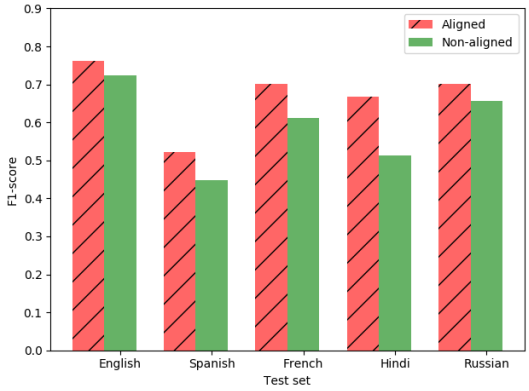
Table 6: (low-resource) Results of low-resource models on the Dutch test set.

For the no-resource scenario, Table 5 provides the results for Dutch and Italian test sets. We hypothesize that language-individual models of languages that have a similar origin as the test language will yield better results than those of languages with a different origin. If we focus on the F_1-scores, we find a general pattern that adheres to this intuition. For the Dutch test set, the English model yields the highest F_1-score. This is to be expected since English and Dutch are the only two West-Germanic languages in this experiment. The French model also performs reasonably well, this can be explained by the fact that French and Dutch are both of European origin. Given that French and Spanish are both Romance languages, we would expect similar results on the Dutch test set. Yet, the Spanish model performs significantly worse and does therefore not follow our intuition. The Russian model also yields worse results than the French and English models. This can be explained by the fact that Russian has a Slavic origin. The Hindi model on the other hand is not able to find any valid relations. Given that all other languages have a European nature and Hindi has an Indo-Iranian nature, this behaviour falls in line with our intuition. A similar pattern can be observed for the Italian test set, albeit less distinct.

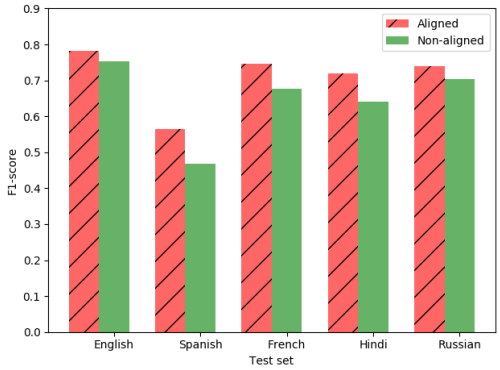
For both Dutch and Italian, the language-consistent model outperforms all language-individual models. This shows the merit of combining languages to find language-consistent relation patterns. In this no-resource scenario, LOREM is a combination of the language-consistent model and the best-performing language-independent model. For the Dutch test set, LOREM even further improves the F_1-score. This is not the case for the Italian test set. These experiments show the first application of an open relation extractor on a different language than it was trained on without the need for a translator. More experiments on different test sets are needed to derive solid conclusions on the matter. Yet, our experiments provide a first indication of the validity of H3.

For the low-resource scenario, if we compare the results shown in the top entry of Table 6 to the evaluation results presented in Table 5, we find that the low-resource Dutch language-individual model is outperformed by the English language-individual model. This indicates that a high-resource model in a similar language outperforms a low-resource model in the test language. However, since we now have a Dutch language-individual model, we can combine it with the language-consistent model to form a full LOREM model. Comparing these results to Table 5, we see that the LOREM model that employs the Dutch language-individual model outperforms all models from the no-resource scenario. This is another indication of the validity of hypothesis 3 for the Dutch test set. Again, more experiments need to be conducted to derive more general conclusions.

Until now, we trained a full language-individual model for the low-resource language, ignoring the fact that we might need to treat a low-resource scenario differently than a high-resource scenario. It is a well-known phenomenon that more complex models generally require more training data, since more parameters need to be optimized. We have examined the possibility of only using a CNN or Bi-LSTM instead of both, to reduce the number of parameters. Results show that although LOREM$_{LSTM}$ and LOREM$_{CNN}$ achieve a higher precision scores than LOREM (0.802 and 0.836 to 0.753), this comes at the expense of a lower recall scores (0.616 and 0.566 to 0.646). As a result, the F_1-scores of are lower than or equal to those of LOREM (0.696 and 0.675 to 0.696). Therefore, we did not find a clear advantage of simplifying the model in this low-resource scenario. Please note that results presented by Jia et al. [12] clearly



(a) F_1-score using aligned and non-aligned embeddings for the language-consistent sub-model.



(b) F_1-score using aligned and non-aligned embeddings for LOREM.

Figure 2: Aligned and non-aligned word embeddings for the language-consistent model and LOREM.

show that combining a CNN and LSTM outperforms both separate models for the high-resource ORE task.
While the main focus of the paper was on high-resource languages, we consider our work to be an initial yet important step towards open relation extraction in no or low-resource languages.

QUALITATIVE ANALYSIS

Next to the quantitative analysis, we also conducted a qualitative analysis on the English test sets. We only performed an English error analysis since it was the only language in which all the authors were fluent.

True positives: We found that LOREM is better at extracting relations that follow abnormal patterns than the language-individual sub-model. For example, given the sentence *"The market wants to do better, said Gregory Bundy, head of equity trading."* and entity tuple `<Gregory Bundy, The market wants to do better>`, the language-individual model does not find a relation, while LOREM extracts `said` as being the relation. Here, we find that the language-consistent component provides additional information which allows relations to be extracted, even if the entities appear in reverse order. It is likely that such patterns occur in multiple languages from which LOREM learned them, even if they were not present in the English training set. Such examples illustrate the benefits of LOREM over a language-individual approach.

False Positives and Negatives: Upon manual inspection, we find that the majority of errors arise from relations that contain multiple words. In these cases, LOREM extracts either too many or too few words compared to the ground truth relations. Typical examples include *"BIC is being sued by people who say their lighters exploded."* and *"The region is still far from rebuilt."*, from which LOREM extracts `is being sued` and `is still`, while the ground truth values are `is being sued by` and `is` respectively. These examples show that although the extraction is not completely correct, the relation is still captured to a certain extent in many cases. The test set also contains sentences from which LOREM can not extract any relations. A typical error occurs when we want to extract relations that occur between more than two entities. Given a sentence like *"28 Square miles of antennae and computers that message smart fridges, robot lawn mowers and smart doorbells vacuum up satellite and radio communications."* with entity tuple `<28 Square miles of antennae, radio communications>`, LOREM finds no relations even though the relation `vacuum up` is present between multiple entities in this sentence.

6 CONCLUSIONS AND FUTURE WORK

In this work, we have presented a Language-consistent Open Relation Extraction Model; LOREM. The core idea is to augment individual open relation extraction mono-lingual models with an additional language-consistent model representing relation patterns shared between languages. Our quantitative and qualitative experiments indicate that harvesting and including such language-consistent patterns improves extraction performances considerably while not relying on any manually-created language-specific external knowledge or NLP tools. Initial experiments show that this effect is particularly valuable when extending to new languages for which no or only little training data is available. In these cases, LOREM and its sub-models can still be used to extract valid relationships by exploiting language consistent relation patterns. As a result, it is relatively easy to extend LOREM to new languages as providing only some training data can be sufficient. However, evaluating with additional languages would be required to better understand or quantify this effect.

Additionally, we conclude that multilingual word embeddings provide an effective approach to introduce latent consistency among input languages, which proved to be beneficial to the performance.

We see many opportunities for future research within this promising domain. More improvements could be made to the CNN and RNN by including more techniques proposed in the closed RE paradigm, such as piecewise max-pooling [22] or varying CNN window sizes [17]. An in-depth analysis of the different layers of these models could shine a better light on which relation patterns are actually learned by the model.

Beyond tuning the architecture of the individual models, enhancements can be made with respect to the language consistent

model. In our current prototype, a single language-consistent model is trained and used in concert with the mono-lingual models we had available. However, natural languages developed historically as language families which can be organized along a language tree (for example, Dutch shares many similarities with both English and German, but of course is more distant to Japanese). Thus, an improved version of LOREM should have multiple language-consistent models for subsets of available languages which indeed posses consistency between them. As a starting point, these could be implemented mirroring the language families identified in linguistic literature, but a more promising approach would be to learn which languages can be effectively combined for boosting extraction performance. Unfortunately, such research is severely hampered by the lack of comparable and reliable publicly available training and especially test datasets for a larger number of languages (note that while the WMORC_auto corpus which we also use covers many languages, it is not sufficiently reliable for this task as it has been automatically generated). This lack of available training and test data also cut short the evaluations of our current variant of LOREM presented in this work. Lastly, given the general set-up of LOREM as a sequence tagging model, we wonder if the model could also be applied to similar language sequence tagging tasks, such as named entity recognition. Thus, the applicability of LOREM to related sequence tasks could be an interesting direction for future work.

REFERENCES

[1] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. 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)*, Vol. 1. 344–354. https://doi.org/10.3115/v1/P15-1034

[2] Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni. 2007. Open information extraction from the web. In *IJCAI*, Vol. 7. 2670–2676. https://doi.org/10.1145/1409360.1409378

[3] Xilun Chen and Claire Cardie. 2018. Unsupervised Multilingual Word Embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 261–270. https://arxiv.org/abs/1808.08933

[4] Lei Cui, Furu Wei, and Ming Zhou. 2018. Neural Open Information Extraction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 407–413. https://www.aclweb.org/anthology/P18-2065

[5] Aron Culotta, Andrew McCallum, and Jonathan Betz. 2006. Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, 296–303. https://doi.org/10.3115/1220835.1220873

[6] Luciano Del Corro and Rainer Gemulla. 2013. Clausie: clause-based open information extraction. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 355–366. https://doi.org/10.1145/2488388.2488420

[7] Anthony Fader, Stephen Soderland, and Oren Etzioni. 2011. Identifying Relations for Open Information Extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1535–1545. https://www.aclweb.org/anthology/D11-1142

[8] Manaal Faruqui and Shankar Kumar. 2015. Multilingual Open Relation Extraction Using Cross-lingual Projection. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 1351–1356. https://doi.org/10.3115/v1/N15-1151

[9] Katrin Fundel, Robert Küffner, and Ralf Zimmer. 2006. RelEx—Relation extraction using dependency parse trees. *Bioinformatics* 23, 3 (2006), 365–371.

[10] Pablo Gamallo and Marcos Garcia. 2015. Multilingual open information extraction. In *Portuguese Conference on Artificial Intelligence*. Springer, 711–722. https://doi.org/10.1007/978-3-319-23485-4_72

[11] Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. https://www.aclweb.org/anthology/L18-1550

[12] Shengbin Jia, Yang Xiang, and Xiaojun Chen. 2018. Supervised Neural Models Revitalize the Open Relation Extraction. *CoRR* abs/1809.09408 (2018). https://arxiv.org/pdf/1809.09408.pdf

[13] Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2979–2984. https://www.aclweb.org/anthology/D18-1330

[14] Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Neural relation extraction with multi-lingual attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 34–43.

[15] Mausam. 2016. Open information extraction systems and downstream applications. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. AAAI Press, 4074–4077. https://www.ijcai.org/Proceedings/16/Papers/604.pdf

[16] Mausam, Michael Schmitz, Robert Bart, Stephen Soderland, and Oren Etzioni. 2012. Open Language Learning for Information Extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 523–534. https://www.aclweb.org/anthology/D12-1048

[17] Thien Huu Nguyen and Ralph Grishman. 2015. Relation Extraction: Perspective from Convolutional Neural Networks. In *Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing*. Association for Computational Linguistics, 39–48. https://doi.org/10.3115/v1/W15-1506

[18] Changqin Quan, Meng Wang, and Fuji Ren. 2014. An unsupervised text mining method for relation extraction from biomedical literature. *PloS one* 9, 7 (2014), e102039.

[19] Gabriel Stanovsky, Julian Michael, Luke Zettlemoyer, and Ido Dagan. 2018. Supervised open information extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Vol. 1. 885–895. https://doi.org/10.18653/v1/N18-1081

[20] Laurent Vannini and Hervé Le Crosnier. 2012. *Net.lang: Towards the Multilingual Cyberspace*. C & F Editions. http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/netlang_EN_pdfedition.pdf

[21] Xiaozhi Wang, Xu Han, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2018. Adversarial Multi-lingual Neural Relation Extraction. In *Proceedings of the 27th International Conference on Computational Linguistics*. 1156–1166. https://www.aclweb.org/anthology/C18-1099

[22] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1753–1762. https://doi.org/10.18653/v1/D15-1203

[23] Sheng Zhang, Kevin Duh, and Benjamin Van Durme. 2017. MT/IE: Cross-lingual open information extraction with neural sequence-to-sequence models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, Vol. 2. 64–70. https://doi.org/10.18653/v1/E17-2011

[24] Sheng Zhang, Kevin Duh, and Benjamin Van Durme. 2017. Selective Decoding for Cross-lingual Open Information Extraction. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1. 832–842. https://www.aclweb.org/anthology/I17-1084

[25] Alisa Zhila and Alexander Gelbukh. 2013. Comparison of open information extraction for English and Spanish. In *19th Annual International Conference Dialog*. 714–722. https://doi.org/10.3115/v1/P14-3011

[26] Alisa Zhila and Alexander Gelbukh. 2016. Open Information Extraction from real Internet texts in Spanish using constraints over part-of-speech sequences: Problems of the method, their causes, and ways for improvement. *Revista signos* 49 (03 2016), 119 – 142. https://scielo.conicyt.cl/pdf/signos/v49n90/a06.pdf