REDFM: a Filtered and Multilingual Relation Extraction Dataset

Pere-Lluís Huguet Cabot
Babelscape, Italy
& Sapienza University of Rome
huguetcabot@babelscape.com

Simone Tedeschi
Babelscape, Italy
& Sapienza University of Rome
tedeschi@babelscape.com

Axel Ngonga
Paderborn University
axel.ngonga@upb.de

Roberto Navigli
Sapienza University of Rome
navigli@diag.uniroma1.it

Abstract

Relation Extraction (RE) is a task that identifies relationships between entities in a text, enabling the acquisition of relational facts and bridging the gap between natural language and structured knowledge. However, current RE models often rely on small datasets with low coverage of relation types, particularly when working with languages other than English.
In this paper, we address the above issue and provide two new resources that enable the training and evaluation of multilingual RE systems. First, we present SREDFM, an automatically annotated dataset covering 18 languages, 400 relation types, 13 entity types, totaling more than 40 million triplet instances. Second, we propose REDFM, a smaller, human-revised dataset for seven languages that allows for the evaluation of multilingual RE systems. To demonstrate the utility of these novel datasets, we experiment with the first end-to-end multilingual RE model, mREBEL, that extracts triplets, including entity types, in multiple languages. We release our resources and model checkpoints at https://www.github.com/babelscape/rebel.

1 Introduction

The vast majority of online and offline content consists of raw, natural language text containing factual information. Current Large Language Models (LLMs) are pretrained on such text, allowing reasoning over it through tasks such as Question Answering (Bouziane et al., 2015) or Text Summarization (El-Kassas et al., 2021). On the other hand, structured resources such as Knowledge Graphs enable knowledge-based, explainable, machine-ready reasoning over their content. Both approaches are important and are widely used within Natural Language Processing systems, with recent trends looking at combining them (Yamada et al., 2020; Sun et al., 2021).

Information Extraction tackles the need for systems that extract structured information from raw text. Specifically, end-to-end Relation Extraction extracts the relational information between entities in a given text, providing a structured prediction. However, although some highly capable systems have been released (Wang and Lu, 2020; Paolini et al., 2021; Huguet Cabot and Navigli, 2021), there are few high-quality, contemporary resources. Current RE datasets are outdated, behind paywalls, have design flaws, or only consider English. While multilingual datasets exist, such as ACE05[1] or SMiLER (Seganti et al., 2021), the former covers only six relation types, and requires a paid license for its use. The latter is more recent, bigger, and has a higher coverage of relation types, but it does not contain humanly-annotated samples to be evaluated reliably and presents issues when framed as an End-to-End Relation Extraction dataset. Instead, the availability of large high-quality resources is fundamental in order to allow LLMs to be trained and evaluated on trustworthy multilingual RE benchmarks.

In this paper we introduce large amounts of high-coverage RE annotated data in a multilingual fashion. Our new resources will enable the training of multilingual RE systems and their evaluation. In particular, we provide three main contributions:

1. We present REDFM, our humanly-revised dataset with 32 relation types and 7 languages;
2. We introduce SREDFM, a silver-standard dataset based on interconnecting Wikipedia and Wikidata, filtered by a Critic system trained on human annotations. It covers 400 relation types, 18 languages, and more than 44M triplet instances. Both datasets are automatically enriched with entity-type information using a novel entity typing approach.
3. We demonstrate the usefulness of these new

[1] https://catalog.ldc.upenn.edu/LDC2006T06

resources by releasing mREBEL, a multilingual system for Relation Classification and Relation Extraction that extracts entity types.

2 Related work

2.1 Relation Extraction

In Relation Extraction (RE), the goal is to identify all triplets, composed of a subject, an object, and a relation between them, within a given text. Early approaches to RE split the task into two different sub-tasks: Named Entity Recognition (NER) (Nadeau and Sekine, 2007), which identifies all entities, and Relation Classification (Bassignana and Plank, 2022), which classifies the relationship, or lack thereof, between them. However, errors from the NER system may be propagated to the subsequent module, leaving the shared information in the interaction of both tasks unexplored.

Recent works have tackled RE in an end-to-end fashion, seeking to overcome these problems by using different abstractions of the task. Miwa and Sasaki (2014) introduced a table representation and reframed RE as a table-filling task. This idea was further explored and extended by Pawar et al. (2017) and Wang and Lu (2020). Nevertheless, these systems still had some restrictions, such as assuming that only one relation exists between each pair. Instead by framing the task as a sequence of triplets to be decoded, seq2seq approaches (Paolini et al., 2021; Huguet Cabot and Navigli, 2021) provided more flexibility to the RE task and lifted some of these restrictions. Nevertheless, seq2seq models are notoriously data-hungry, hence vast amounts of data are needed to enable them to learn the task with satisfactory scores.

2.2 Relation Extraction Datasets

Datasets play a crucial role in training and evaluating RE systems. However, manually annotating RE data is a costly and time-consuming process. As a result, many RE datasets have been created using distant supervision methods, such as NYT (Riedel et al., 2010), T-REx (Elsahar et al., 2018) or DocRED (Yao et al., 2019). Despite their widespread use in the RE community, these datasets have limitations. For instance, automatically generated datasets can contain noisy labels, leading to unfair or misleading evaluations. Additionally, there has been a long-standing focus on monolingual relation extraction systems, particularly in English.

The ACE05 benchmark presented some of the first relation extraction datasets in three languages, Arabic, Chinese, and English, but the focus on Arabic and Chinese quickly faded away while resources for English continued to grow. One of the main challenges in developing multilingual relation extraction systems is the lack of annotated data for the task. The SMiLER dataset (Seganti et al., 2021), based on distant supervision, uses Wikipedia and Wikidata to create a multilingual relation extraction dataset. However, besides being automatic, SMiLER limits annotations to one triple per instance. To improve the quality of multilingual data, in this paper we present a new resource that, despite being based on automatic annotation, includes manual annotation and enables wide-coverage higher-quality RE, thus overcoming the limitations of the existing datasets.

3 RED^{FM}

In this Section, we present RED^{FM}, our supervised and multilingual dataset for Relation Extraction, and a larger SRED^{FM}, a silver-annotated dataset covering more languages and relation types. The creation of the dataset consists of several steps: data collection and processing (Section 3.1), manual annotation (Section 3.2), a triplet filtering system (Section 3.3) and entity typing (Section 3.4). Figure 1 shows an overview of this process.

3.1 Data Extraction

We base our dataset on Wikidata and Wikipedia, and expand cRocoDiLe, the data extraction pipeline from Huguet Cabot and Navigli (2021), to obtain a large collection of triplets in multiple languages (see Appendix A for more details). We use the hyperlinks from Wikipedia abstracts, i.e. the content before the Table of Contents, as entity mentions and the relations in Wikidata between them. We run our pipeline in the following 18 languages: Arabic, Catalan, Chinese, Dutch, German, Greek, English, French, Hindi, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Swedish, and Vietnamese. Then, we collapse inverse relations and keep the 400 most frequent ones. We highlight that some extracted relations are not necessarily entailed by the Wikipedia text; therefore, we apply a multilingual NLI system[2] to filter out those with a low entailment score (i.e. < 0.1).

Despite using NLI techniques to filter out false positives, distant RE annotations still present noisy

[2] xlm-roberta-large-xnli

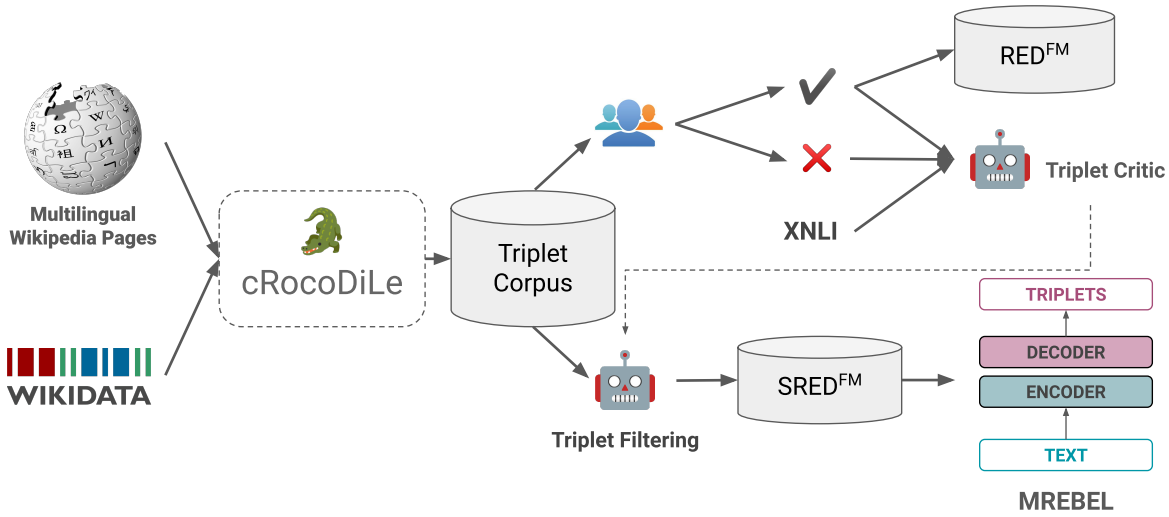

Figure 1: Our full pipeline for the creation of RED^{FM}, SRED^{FM} and mREBEL.

labels. This can result in unfair or misleading evaluations, as demonstrated for TACRED (Zhang et al., 2017), which had 23.9% wrongly annotated triplets that were later revised and corrected by Stoica et al. (2021). Moreover, our triplet corpus extraction pipeline relies on existing triplets in Wikidata, similar to T-REx (Elsahar et al., 2018). These latter showed how certain relation types, such as "capital", have a lower entailment score and may not be entailed by a given text even though the entities involved share the relation in Wikidata.

Given these challenges in distant RE annotation, manual filtering of a portion of the data is necessary to ensure high-quality, accurate annotations.

3.2 Manual Annotation

We manually filter a portion of the data to deal with false positives present in the dataset for a subset of languages (i.e. Arabic, Chinese, German, English, French, Italian, and Spanish) through crowdsourced annotation:

1. We reduce the coverage of the annotated data to the top 32 most frequent relation types. See Appendix B for details on each of these types.

2. We select a portion of our silver annotated data consisting of i) common Wikipedia pages across those languages and ii) a random sample with less frequent relations to balance the dataset.

3. We ask human annotators to validate each triplet. They are shown the context text with subject and object entities highlighted, and the possible relation between them from the silver extraction. They must answer whether the text conveys the necessary information to infer that the relationship between those two entities is true.

4. We annotate each triple three times using different annotators, obtaining an average inter-rater reliability (Krippendorff's alpha) across languages of $\alpha_\kappa = 0.73$.

5. We keep as true positives those relations with at least two annotators answering true. We consider the rest false positives.

	ar	de	fr	en	es	it	zh
Annotators	3	9	12	9	10	13	7
α_κ	0.76	0.80	0.77	0.72	0.61	0.73	0.70
Filtered (%)	6.79	5.47	4.40	8.54	12.53	8.93	12.11

Table 1: Number of annotators, their agreement based on Krippendorff's alpha, and the percentage of filtered triplets for each language.

We employed Amazon Mechanical Turk and manually selected annotators that qualified for the task in each language. The annotation scheme can be found in Appendix B. From Table 1 we see that around 8% of annotated triplets, on average, were labeled as non-entailed by the context provided, albeit the percentage varies across languages. For instance, Spanish had a lower agreement across annotators and a higher number of filtered instances.

		All True	XNLI	Ours	+XNLI$^-$	+XNLI
Arabic	**R.**	100.0	76.6	96.9	**98.6**	98.5
	P.	93.2	**94.9**	**94.9**	94.5	94.5
	F1	96.5	84.8	95.9	**96.5**	**96.5**
	Acc.	93.2	74.3	92.4	**93.3**	**93.3**
Chinese	**R.**	100.0	82.3	98.7	98.8	**99.6**
	P.	87.9	89.9	**90.1**	**90.1**	89.7
	F1	93.5	85.9	94.2	94.2	**94.4**
	Acc.	87.9	76.3	89.3	89.4	**89.5**
Total	**R.**	100.0	79.1	97.7	98.7	**98.9**
	P.	90.8	92.6	**92.8**	92.6	92.4
	F1	95.2	85.3	95.2	**95.5**	**95.5**
	Acc.	90.8	75.2	91.0	**91.6**	**91.6**

Table 2: Performance for the Triplet Critic. **All True** shows baseline when all triplets are marked as True. **XNLI** uses the entailment prediction of a model trained solely on XNLI. **Ours** is trained solely on our annotated data without ar/zh. Last two columns show the multi-task approach with (**+XNLI**) and without ar/zh (**+XNLI**$^-$) data from XNLI.

3.3 Triplet Critic

Our manual annotation procedure (Section 3.2) filtered a portion of the silver data in order to have a higher-quality subset on which to train and evaluate our models. However, by removing the negative triplets we disregard valuable information that can be used to improve the quality of the remaining annotations, i.e. all those not validated by humans. Inspired by West et al. (2021), who trained *critics* based on human annotations on commonsense triplets, we use our annotated triplets, both true and false positives with their contexts, to train a Triplet Critic. Specifically, given a textual context c and an annotated triplet t that may appear in c, we train a cross-encoder $T(c, t)$ to predict whether c, the premise, entails t, the hypothesis. This setup was inspired by NLI and results in training examples such as:

Premise	Hypothesis	Label
Telefe (acronym for Televisión Federal) is a television station located in Buenos Aires, Argentina.	Telefe instance of television station	True
	Buenos Aires country Argentina	True
	Argentina capital Buenos Aires	False

Once T is trained, we can use it on our silver data to filter out other false positives, i.e., triplets that, albeit present in Wikidata for two entities within the context, are not entailed by that context.

We test our approach by training T in English, French, Italian, Spanish and German, namely European languages with shared families (Romance and Germanic), and testing on Arabic and Chinese. This setup will test the zero-shot multilingual capabilities on unseen languages in order to determine whether the Critic can be applied to any language. We base our Triplet Critic on DeBERTaV3 (He et al., 2021) with a classification head on top of the [CLS] token that produces a binary prediction, trained using a Cross-Entropy loss criterion. Furthermore, since the task is similar to and inspired by NLI, we explore a multi-task approach using the XNLI dataset (Conneau et al., 2018), aiming at improving cross-lingual performance. To this end, we add an additional linear layer at the end of the model for NLI that projects the output layer to the three possible predictions (neutral, contradiction, entailment), again using a Cross-Entropy loss.

Table 2 shows the Triplet Critic results when trained under different setups. We see how the use of our data dramatically improves upon the XNLI baseline, especially in terms of accuracy. While we are primarily interested in precision so as to guarantee that triplets are valid, a low accuracy would lead to missing annotations, which we also want to avoid. Additionally, when our Triplet Critic is trained simultaneously on our data and XNLI, even if no Arabic or Chinese data is used (i.e. +XNLI$^-$), performance further improves: the system achieves an average 92.6% precision, on par with using only our data, but sees a point increase in recall, which is remarkable, taking into account the high class inbalance. Instead, when XNLI data from those languages is added (i.e. +XNLI), we observe a small trade-off between precision and recall.

Overall, these zero-shot results certainly legitimize the use of our Triplet Critic to refine our silver data for unseen languages, and suggest even more promising benefits for seen languages. Furthermore, the Triplet Critic serves as a feedback on the consistency of human annotations since the models have successfully learned from them.

3.4 Entity Typing

In RE datasets, the entity types are commonly included in the triplets (Riedel et al., 2010) and, therefore, need to be predicted by the models under the strict evaluation as described in Taillé et al. (2020). This Section describes the procedure through which we automatically label entities with their types.

We start by mapping entities in Wikipedia to Ba-

belNet (Navigli and Ponzetto, 2012; Navigli et al., 2021) synsets by exploiting the one-to-one linkage between them. Then, we annotate synsets by applying the knowledge-based semantic classifier introduced by Tedeschi et al. (2021b), which exploits the relational information in BabelNet such as hypernymy and hyponymy relations. This procedure yields ∼7.5M entities labeled with an entity type. However, since the annotations are automatically derived and prone to errors, we devise a new strategy to improve their quality. Specifically, we design a Transformer-based classifier that takes a synset and returns its NER category. More formally, let us define the functions $L(s)$ and $D(s)$ that output the main lemma and the textual description of a synset s, respectively. Then, given a synset s and the above-defined functions, we provide the string $I(s, D, L)$ = [CLS] $L(s)$ [SEP] $D(s)$ [SEP] as input to the classifier that predicts a label e ∈ E[3]. The tagset E is obtained by refining the categorization of named entities introduced by Tedeschi et al. (2021a) based on the ability of automated systems to distinguish NER categories and on the frequency of these categories in Wikipedia articles (Tedeschi and Navigli, 2022). To train the classifier, we construct a dataset by selecting a high-quality subset from the 7.5M automatically-produced annotations by taking only synsets with a maximum distance equal to 1 from one of the 40k synsets in WordNet (Miller, 1995), this latter being a manually-curated subset of BabelNet. By doing this, we obtain a set consisting of 1.2M high-quality annotations that we split into 80% for training and 20% for validation, and convert these to the above-specified $I(s, D, L)$ format.

Finally, we use the trained classifier to confirm or replace the previous 7.5M annotations, resulting in 6.2M (82.4%) confirmations and 1.3M (17.6%) changes, and employ it to label new Wikidata instances as well, thus obtaining a final mapping consisting of ∼13M Wikidata entries annotated with their entity types. By manually inspecting a sample of 100 changes, we observed that our NER classifier was right 68% of the time, wrong 17%, while the remaining 15% of the time both annotations were wrong, providing an improvement of 51% over 1.3M changes. We highlight that 68% is not the accuracy of our classifier as it is computed on 100 items where there is a disagreement between the original annotations produced by WikiNEuRal (Tedeschi et al., 2021b) – the current state of the art in entity typing – and the annotations produced by our model. Indeed, an accuracy of 68% on this subset means that our classifier corrected most of the instances that were previously mistaken by WikiNEuRal. For completeness, we report that when the two systems agree, i.e. 82% of the times, they are correct in 98% of these cases, as measured on another subset of 100 instances.

[3]E = {location, person, number, time, organization, date, event, celestial body, media, disease, concept, miscellaneous and unknown}

	Human Annotated				Distant supervision			
	ACE05	CONLL04	DocRED	RED^{FM}	DocRED	NYT	SMILER	SRED^{FM}
Docs.	1.6K	1.4K	5K	15.4K	100K	66.2K	1.1M	12.3M
Sents.	30.9K	1.4K	-	43.7K	-	66.2K	1.1M	46.6M
Rels.	6	5	96	32	96	24	36	400
Ents.	5	4	6	13	6	3	-	13
AR	4.7K	-	-	1.8K	-	-	9K	3.3M
CA	-	-	-	-	-	-	-	1.7M
DE	-	-	-	7.5K	-	-	53K	4.8M
EL	-	-	-	-	-	-	-	325K
EN	8.7K	6.8K	58.6K	10.9K	1M	111K	748K	12.4M
ES	-	-	-	6.5K	-	-	12K	4.2M
FA	-	-	-	-	-	-	3K	-
FR	-	-	-	7.4K	-	-	62K	4.2M
HI	-	-	-	-	-	-	-	301K
IT	-	-	-	6.8K	-	-	76K	2M
JA	-	-	-	-	-	-	-	3.3M
KO	-	-	-	-	-	-	20K	1M
NL	-	-	-	-	-	-	40K	3M
PL	-	-	-	-	-	-	17K	3.7M
PT	-	-	-	-	-	-	45K	2.7M
RU	-	-	-	-	-	-	7K	1.6M
SV	-	-	-	-	-	-	5K	7.2M
UK	-	-	-	-	-	-	1K	-
VI	-	-	-	-	-	-	-	1.4M
ZH	9.3K	-	-	1.4K	-	-	-	3M

Table 3: Number of relation types (Rels.), entity types (Ents.) and annotated triplets in RE resources.

3.5 SRED^{FM}

The current datasets for Relation Extraction often lack complete coverage of relations. The SMiLER dataset (Seganti et al., 2021) only annotates one triplet per instance, resulting in a limited understanding of the relationships therein. For example, in the instance "*Fredrik Hermansson (born 18 July 1976) is a Swedish musician. He was a keyboardist and backing vocalist in the Swedish progressive rock band Pain of Salvation.*", the triplet (Fredrik Hermansson, has-genre, progressive rock) is annotated, but other triplets such as (Fredrik Hermansson, has-occupation, musician) and (Fredrik Hermansson, has-nationality, Swedish) are also valid.

Another common issue with RE datasets is the high class imbalance, particularly for distantly annotated datasets. This is often due to the skewed distributions that are intrinsic in knowledge bases

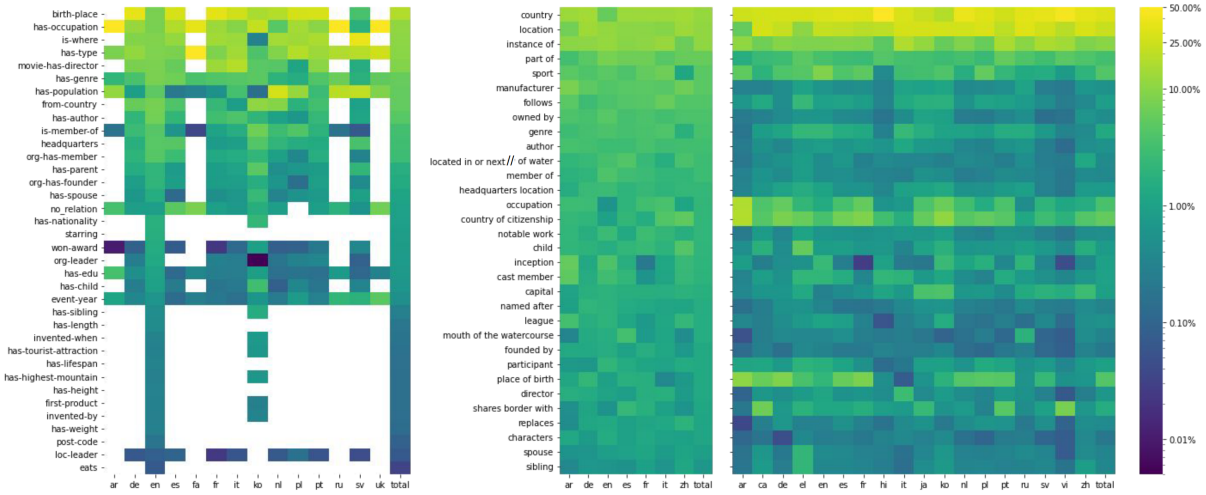

Figure 2: Comparison of relation type distribution by percentage between SMiLER (left), RED^{FM} and SRED^{FM} (right). Best seen in color.

such as Wikidata, which are often used to construct RE resources. This leads to low coverage for lower frequency classes, as seen in Figure 2 for certain languages in the SMiLER dataset. In DocRED (Yao et al., 2019), another distantly annotated dataset, location-based relations constitute over 50% of instances.

Fully human-annotated datasets that overcome these limitations are scarce and often not widely accessible. Additionally, they often cover a narrow set of languages and relations (Table 3). To address these issues, we introduce Silver RED^{FM}. SRED^{FM} is a large, multilingual RE dataset that contains more than 45M triplets and covers 400 relation types and 18 languages. It is created using the data extraction procedure described in Section 3.1 and the Triplet Critic introduced in Section 3.3. SRED^{FM} overcomes some of the previous shortcomings of current datasets by providing a higher coverage of annotation and more evenly distributed classes.

Using the same example sentence, in SRED^{FM}, the following triplets are annotated: (Fredrik Hermansson, country of citizenship, Swedish), (Fredrik Hermansson, occupation, musician), (Fredrik Hermansson, date of birth, 18 July 1976), and (Fredrik Hermansson, member of, Pain of Salvation). Regarding class balance, in the English portion of SRED^{FM} location-based relations make up less than 37%. Hence our datasets have more evenly distributed classes. Figure 2 (right) shows the distribution for the top 32 of the 400 relation types in SRED^{FM}.

Additionally, we provide a pipeline that enables the automatic creation of an RE dataset in any language. So, even though we release the SRED^{FM} dataset as described in this paper (i.e. covering 18 different languages), we encourage the expansion to other languages by using our pipeline available here.

In summary, SRED^{FM} is a large, multilingual dataset that addresses the shortcomings of current datasets by providing a higher coverage of annotation and more evenly distributed classes. RED^{FM}, instead, is the result of the manual annotation (Section 3.1) to which we add entity types. We split them into training, validation and test, with no overlapping Wikipedia pages across splits. Details can be found in Table 10 in Appendix E.

4 mREBEL

In this section, we present our system, mREBEL (Multilingual Relation Extraction By End-to-end Language generation), which is a multilingual relation extraction model pre-trained on SRED^{FM}. It is a multilingual extension of the REBEL model introduced in Huguet Cabot and Navigli (2021), which uses a seq2seq architecture to convert relations into text sequences that can be decoded by the model. We convert triplets into text sequences and pre-train our model using mBART-50 (Tang et al., 2021). To support multiple languages, we prepend the input text with a special language token (i.e. `en_XX`). Additionally, we include relation classification (RC) in the pre-training phase of mREBEL. Specifically, for 5% of the training data, we select a random triplet, mark the subject and object entities in the input text, and use a special token

		Input text	Triplets	Linearized Triplets
Classification	nl	nl_XX # Mumbai Mirror # is een Engelstalige tabloid, die verschijnt in de Indiase stad Mumbai. Het is hier met een oplage van zo'n 700.000 exemplaren de belangrijkste krant. Het dagblad verscheen voor het eerst op @ 30 mei 2005 @,	(Mumbai Mirror, inception, 30 mei 2005)	tp_XX<relation>Mumbai Mirror <media>30 mei 2005 <date>inception
Extraction	ca	ca_XX Can Verboom és una masia amb elements gòtics i barrocs de Premià de Dalt (Maresme) protegida com a bé cultural d'interès local.	(Can Verboom, located in the administrative territorial entity, Premià de Dalt) (Can Verboom, heritage designation, bé cultural d'interès local)	tp_XX<triplet>Can Verboom <loc> Premià de Dalt <loc>located in the administrative territorial entity <loc> bé cultural d'interès local <loc> heritage designation

Table 4: Examples from SRED^{FM} with their triplet linearization used to train mREBEL.

	ar	de	en	es	fa	fr	it	ko	nl	pl	pt	ru	sv	uk	Avg.
Support	190	1049	731	225	54	1241	1501	228	791	343	880	131	92	20	
HERBERTa	49.0	63.0	77.0	46.0	72.0	58.0	69.0	51.0	61.0	50.0	62.0	30.0	59.0	25.0	54.7
mREBEL$^{T}_{400}$	**75.3**	63.1	**77.7**	**57.8**	69.2	64.3	**83.5**	**62.5**	72.8	76.1	69.8	71.0	76.1	**70.0**	70.5
mREBEL$_{400}$	73.7	**64.2**	77.7	54.7	**74.8**	**66.1**	82.9	61.7	**73.6**	**76.4**	**70.1**	**75.6**	**78.3**	65.0	**70.8**
mT5$_{BASE}$*	95.1	95.4	96.1	81.1	73.1	97.2	98.3	83.2	96.9	95.6	96.9	87.6	63.0	71.8	88.0
mT5$_{BASE}$(en)	94.0	94.9	-	91.7	91.1	96.0	97.5	78.2	97.5	93.3	95.2	93.8	97.8	94.7	93.5
mREBEL$^{T}_{B400}$	**99.5**	96.8	96.6	95.1	**100.0**	97.7	98.7	**94.7**	98.3	97.4	97.8	**100.0**	96.7	**100.0**	97.8
mREBEL$^{T}_{400}$	**99.5**	97.4	97.5	94.9	97.0	**97.8**	98.8	93.1	**98.7**	98.4	98.3	**100.0**	97.8	**100.0**	97.8
mREBEL$^{T}_{400}$*	**99.5**	96.9	97.5	93.5	99.0	**97.8**	**98.9**	94.1	98.0	98.3	97.7	98.8	**98.4**	97.4	97.6
mREBEL$_{400}$	**99.5**	**97.5**	**97.7**	**95.3**	**100.0**	97.6	**98.9**	**94.7**	98.6	**98.7**	**98.4**	99.2	97.8	**100.0**	**98.1**

Table 5: Results on SMiLER. Micro-F1 scores per language. Top half shows RE, bottom half RC. Selected top performing multilingual HERBERTa per language from (Seganti et al., 2021) and mT5$_{BASE}$ from Chen et al. (2022). Chen et al. (2022)(en) was trained on English data. * indicates separate training per language.

<relation> to indicate to the model that only one triplet needs to be decoded. Finally, to promote cross-lingual transfer, we use the English names for the relation types when decoding the triplets. Table 4 shows how instances from SRED^{FM} are used to train mREBEL.
We train three versions of mREBEL:

1. mREBEL$^{T}_{400}$, trained on 400 relation types, including entity types;

2. mREBEL$^{T}_{32}$, fine-tuned on top of the previous one but including only the 32 relation types from RED^{FM};

3. mREBEL$^{T}_{B400}$, trained on top of M2M100 (Fan et al., 2020).

For (1) and (2) we also train their untyped versions, mREBEL$_{400}$ and mREBEL$_{32}$.

5 Experimental Setup

We evaluate mREBEL and its variants on our own datasets, i.e. RED^{FM} and SRED^{FM}, and on SMiLER (Seganti et al., 2021). Unless stated, we train on the training sets of all languages simultaneously and apply early stopping based on the Micro-F1 obtained on the overall validation set. We use the Adafactor optimizer and the same Cross-Entropy loss with teacher forcing from Huguet Cabot and Navigli (2021). The full list of hyperparameters is detailed in Appendix D.

Multilingual Relation Extraction We report the Micro-F1 score per language for both SMiLER and RED^{FM} test sets. When evaluating on SMiLER, we use mREBEL$_{400}$ variants as starting checkpoints and fine-tune them on SMiLER training sets. For RED^{FM}, instead, we include the mREBEL$_{32}$ model in our experiments as it was trained on the same set of relations. The inclusion of this model lets us analyze the impact of further fine-tuning on RED^{FM} gold data against the quality of our silver annotation process. As an extrinsic evaluation of our Triplet Critic model from Section 3.3, we train a version of mREBEL$^{T}_{32}$ without filtering triplets.

Multilingual Relation Classification Even though Seganti et al. (2021) introduced SMiLER as a RE dataset, each sentence contains just one

Model	Fine-tuning	ar	de	en	es	fr	it	zh	P	R	F1
mREBEL$^{T\dagger}_{32}$	✘	43.4	53.9	50.0	46.9	48.3	56.4	38.5	43.1	56.1	48.7
mREBEL$^{T}_{32}$	✘	**45.5**	57.8	53.3	**51.4**	**52.5**	**57.8**	**41.8**	47.6	**57.3**	**52.0**
mBART	✔	16.1	39.6	32.6	27.3	28.5	30.0	0.0	26.8	29.0	27.9
mREBEL$^{T}_{B400}$	✔	39.9	50.3	49.0	41.1	41.7	50.6	38.0	45.0	45.1	45.1
mREBEL$_{400}$	✔	33.2	50.0	42.2	38.4	40.3	49.2	30.7	40.9	41.7	41.3
mREBEL$^{T}_{400}$	✔	39.3	52.8	49.5	45.9	46.8	54.7	35.2	47.5	47.0	47.2
mREBEL$^{T\dagger}_{32}$	✔	43.7	55.4	**54.0**	46.9	50.5	57.1	38.8	47.9	53.0	50.3
mREBEL$^{T}_{32}$	✔	43.8	**58.3**	53.7	50.1	51.8	**57.8**	41.3	**48.9**	54.7	51.6

Table 6: Results on RED$^{\text{FM}}$ test set. Micro-F1 scores per language. † indicates the Critic was not used to filter pre-training data. Fine-tuning indicates fine-tuning on RED$^{\text{FM}}$ training set.

	ar	ca	de	el	en	es	fr	hi	it	ja	ko	nl	pl	pt	ru	sv	vi	zh	all
Sents. (K)	4.5	2.7	5.7	0.9	6.9	8.8	4.0	0.4	2.2	3.4	0.8	2.8	6.5	4.5	2.3	7.6	2.0	2.7	68.6
Trip. (K)	27.9	15.5	32.1	4.9	40.0	54.9	26.1	2.8	12.8	29.4	4.4	16.4	36.6	27.5	12.9	44.9	10.5	27.4	427.0
Precision	61.1	57.0	58.5	46.1	62.3	60.6	60.0	50.2	57.7	48.6	48.5	57.5	64.6	60.0	53.4	63.2	59.0	41.0	58.5
Recall	48.7	45.9	44.5	34.1	47.2	51.2	45.9	29.6	44.9	44.2	37.7	49.0	55.4	48.9	38.2	55.7	45.3	36.6	47.9
Micro-F1	54.2	50.8	50.6	39.2	53.7	55.5	52.0	37.2	50.5	46.3	42.5	52.9	59.6	53.9	44.6	59.2	51.2	38.7	52.7
Macro-F1	24.0	24.8	29.3	12.5	36.3	30.5	29.1	8.3	27.9	25.1	17.3	27.6	30.5	29.5	23.5	30.1	19.5	20.3	24.8

Table 7: Results for SRED$^{\text{FM}}$ test set with mREBEL$^{T}_{400}$ on 400 relation types.

annotated triplet and includes the "no relation" class as part of its annotation scheme. Therefore, it is better approached as an RC task and it is more akin to a dataset like TACRED. For instance, Chen et al. (2022) use it as an RC dataset with an array of prompt-based approaches, and we compare our approach with theirs for RC.

6 Results

Multilingual Relation Extraction First, in Table 5 we show how our system performs compared to HERBERTa, the system proposed by Seganti et al. (2021) for SMiLER, using their best-performing setup for each language. We consider this dataset better suited for RC. However, as it originally reports on RE, we demonstrate how our system can perform better when pretrained on SRED$^{\text{FM}}$. In particular, mREBEL$^{T}_{400}$ provides an improvement of about 15 Micro-F1 points compared to HERBERTa. Additionally, as SMiLER does not include entity types, we observe that mREBEL$_{400}$ performs marginally better than mREBEL$^{T}_{400}$.

Table 6, instead, shows the results on RED$^{\text{FM}}$, compared against an mBART baseline. Specifically, we analyze model performance when fine-tuning is, or is not, performed on the train set of RED$^{\text{FM}}$. While performances vary across languages, the best overall Micro-F1 (52.0) is obtained when training on SRED$^{\text{FM}}$, mREBEL$^{T}_{32}$, without further fine-tuning. This confirms that our silver annotation procedure produces high-quality data, as there is no need for further tuning with RED$^{\text{FM}}$, which achieved 51.6. We also see how filtering by the Triplet Critic was crucial: when removed, performance dropped by more than 3 points.

Training on 400 relation types does lead to lower results, since there is a mismatch between the two stages of training. However, mREBEL$^{T}_{400}$ showed decent performance on SRED$^{\text{FM}}$ as shown in Table 7. This provides the first RE system to competitively extract up to 400 relation types in multiple languages. See Appendix C for more results.

Multilingual Relation Classification From the bottom half of Table 5, we can observe how our mREBEL models consistently outperform competitive baselines, i.e. mT5$_{BASE}$* and mT5$_{BASE}$(en), by a large margin on all tested languages.

6.1 Error Analysis

In order to understand the sources of error when training on the RED$^{\text{FM}}$ dataset, we performed an error analysis with mREBEL$^{T}_{32}$. Our study revealed that 27.8% of the errors in the test set can be attributed to specific reasons.

First, there were discrepancies between predicted entity types and those annotated (7.2%). These errors may have arisen from the automatic

Error Type	Count	Example
Entity type	262 (7.2%)	The Peugeot 408 (CONCEPT, MISC) is a small family car produced by Peugeot (ORG) . [manufacturer]
Span underlap	157 (4.3%)	La Changhe Aircraft Industries Corporation (ORG) è un'azienda cinese, specializzata nella produzione di elicotteri (CONCEPT) . [manufacturer]
Span overlap	138 (3.8%)	La caja de Pandora (MEDIA), película muda (CONCEPT) alemana, dirigida por G. W. Pabst. [genre]
Subject	269 (7.4%)	L'œuvre contient notamment " Le Pré Béjine (MEDIA) ", dont le titre a été repris par Sergueï Eisenstein (PERSON) pour un film inachevé (MEDIA) . [director] [director]
Object	117 (3.2%)	OpenSolaris (MISC) ist ein auf Unix (CONCEPT) basierendes Betriebssystem (CONCEPT) für die Plattformen PC, SPARC und andere. [instance of] [instance of]
Relation	66 (1.8%)	The Red Hot Chili Peppers (ORG) were formed in Los Angeles by Kiedis, Flea, guitarist Hillel Slovak (PER) and drummer Jack Irons. [member of] [part of]
Other	2606 (72.2%)	René L'Hermitte (PER) est un journaliste (CONCEPT) et un professeur des universités français, membre de la Résistance et du Parti communiste [occupation]

Table 8: Types of error encountered in RED^{FM}. Orange shows the mismatched prediction, and green is the annotated counterpart. Best seen in color.

nature of the typing annotation, errors by the system or ambiguity in some cases, such as fictional characters, which can be considered either PERSON or MEDIA. Additionally, a portion of errors (8.1%) resulted from mismatches between the predicted and annotated spans for each entity, which may also be ambiguous (see the Span overlap example in Table 8). Another 10.6% of errors were caused by either the subject or object entity being completely misaligned with the annotation. We identify some of these as co-reference errors, such as the example for Subject in Table 8. Evaluation for RE systems often ignores other mentions for the same entity. We believe co-reference resolution has not been properly explored within RE evaluation and this may open interesting opportunities for future work.

Finally, it is worth noting that only 1.8% of errors were due to the wrong relation type being predicted between entities. We consider this to be a strong indicator on the quality of annotated relations between entities. However, we also observed that 72.2% of the errors were caused by incorrect predictions or missing annotations, highlighting the main shortcoming of our annotation procedure. Our approach is based on annotated hyperlinks in Wikipedia and relations in Wikidata, which can result in recall issues where entities in the text are not identified as hyperlinks or relational facts are not present in Wikidata.

7 Conclusions

In this paper, we have addressed some of the key issues facing current multilingual relation extraction datasets by presenting two new resources: SRED^{FM} and RED^{FM}. SRED^{FM} is an automatically annotated dataset covering 18 languages, 400 relation types, 13 entity types, and more than 40 million triplet instances, while RED^{FM} is a smaller, humanly-revised dataset for seven languages. We improved the quality of the entity type annotations in these datasets by using a Transformer-based NER classifier. We also introduced the Triplet Critic, a cross-encoder that is trained on annotated data to predict whether a given context entails a triplet. We demonstrated the utility of these new resources by training new, capable multilingual relation extraction models and evaluating them using our supervised data. We also presented mREBEL, the first multilingual end-to-end relation extraction system that extracts triplets, including entity types. Our work thus contributes to the development of better multilingual relation extraction systems and provides valuable resources for future research.

8 Limitations

There are several limitations to the work presented in this paper that need to be acknowledged.

First, the SRED^{FM} and RED^{FM} datasets are based on Wikipedia and Wikidata, which means they may not cover all possible relation types or entities. In addition, the quality of the annotations in these datasets may be influenced by the biases and limitations of these sources.

Second, the Triplet Critic is trained on a small subset of the SRED^{FM} dataset, which may limit its ability to generalize to other relation types or languages. Additionally, the performance of the Triplet Critic may be affected by the quality of the annotations used to train it.

Third, the authors of this work are native speakers of some of the languages tackled in this work and external native speakers created the annotation guidelines. However, for some of the automatically-annotated languages, there were no native speakers involved. Additionally, the qualitative error analysis does not include Arabic or Chinese examples, as neither of the authors of the paper is proficient in those languages.

Finally, the mREBEL system is based on a Transformer architecture, which may not be optimal for all relation extraction tasks. It is possible that other types of model, such as graph neural networks or rule-based systems, could outperform mREBEL on certain relation types or languages.

Overall, the results presented in this paper should be interpreted in the context of these limitations. Further research is needed to address these limitations and to improve the performance of multilingual relation extraction systems.

9 Ethics Statement

In this work, we present two new relation extraction datasets, RED^{FM} and SRED^{FM}, which are created using distant supervision techniques and the use of human annotation to filter out false positives. We believe that our datasets will help advance the field of relation extraction by providing a high-quality multilingual resource for researchers and practitioners.

We take the ethical considerations of our work seriously. The annotation of the RED^{FM} dataset is based on existing triplets in Wikidata, which may not always reflect the true relation between entities in a given text. Moreover, the use of human annotation ensures a higher level of accuracy in our dataset, but it also raises ethical considerations. We recognize that human annotation may contain errors or biases. Therefore, we encourage researchers to use our dataset with caution and to perform thorough evaluations of their methods. Additionally, we are transparent about our annotation costs and payment to human annotators.

In conclusion, we believe that our dataset and the research it enables will contribute positively to the field of relation extraction, but we also acknowledge that there are ethical considerations that need to be taken into account when using it.

Acknowledgments

The authors gratefully acknowledge the support of the European Union's Horizon 2020 research project *Knowledge Graphs at Scale* (KnowGraphs) under the Marie Skłodowska-Curie grant agreement No. 860801.

This research has been carried out while Pere-Lluís Huguet Cabot and Simone Tedeschi were enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome and with the support of the PNRR MUR project PE0000013-FAIR.

We sincerely thank all annotators who took part in the task as this work would not have been possible without their contribution.

References

Elisa Bassignana and Barbara Plank. 2022. What do you mean by relation extraction? a survey on datasets and study on scientific relation classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 67–83, Dublin, Ireland. Association for Computational Linguistics.

Abdelghani Bouziane, Djelloul Bouchiha, Noureddine Doumi, and Mimoun Malki. 2015. Question answering systems: Survey and trends. *Procedia Computer Science*, 73:366–375. International Conference on Advanced Wireless Information and Communication Technologies (AWICT 2015).

Yuxuan Chen, David Harbecke, and Leonhard Hennig. 2022. Multilingual relation classification via efficient and effective prompting. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Online and Abu Dhabi, the United Arab Emirates. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679.

Hady Elsahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon Hare, Frederique Laforest, and Elena Simperl. 2018. T-REx: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.

Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2370–2381, Punta Cana, Dominican Republic. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Makoto Miwa and Yutaka Sasaki. 2014. Modeling joint entity and relation extraction with table representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1858–1869, Doha, Qatar. Association for Computational Linguistics.

David Nadeau and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Roberto Navigli, Michele Bevilacqua, Simone Conia, Dario Montagnini, and Francesco Cecconi. 2021. Ten years of BabelNet: A survey. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4559–4567.

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cicero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021*.

Sachin Pawar, Pushpak Bhattacharyya, and Girish Palshikar. 2017. End-to-end relation extraction using neural networks and Markov Logic Networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 818–827, Valencia, Spain. Association for Computational Linguistics.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases*, pages 148–163, Berlin, Heidelberg. Springer Berlin Heidelberg.

Alessandro Seganti, Klaudia Firląg, Helena Skowronska, Michał Satława, and Piotr Andruszkiewicz. 2021. Multilingual entity and relation extraction dataset and model. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1946–1955, Online. Association for Computational Linguistics.

George Stoica, Emmanouil Antonios Platanios, and Barnabas Poczos. 2021. Re-tacred: Addressing shortcomings of the tacred dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13843–13850.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation.

Bruno Taillé, Vincent Guigue, Geoffrey Scoutheeten, and Patrick Gallinari. 2020. Let's Stop Incorrect Comparisons in End-to-end Relation Extraction! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3689–3701, Online. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. Multilingual translation from denoising pre-training. In *Findings of the Association*

for Computational Linguistics: ACL-IJCNLP 2021, pages 3450–3466, Online. Association for Computational Linguistics.

Simone Tedeschi, Simone Conia, Francesco Cecconi, and Roberto Navigli. 2021a. Named Entity Recognition for Entity Linking: What works and what's next. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2584–2596, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021b. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Simone Tedeschi and Roberto Navigli. 2022. MultiNERD: A multilingual, multi-genre and fine-grained dataset for named entity recognition (and disambiguation). In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 801–812, Seattle, United States. Association for Computational Linguistics.

Jue Wang and Wei Lu. 2020. Two are better than one: Joint entity and relation extraction with table-sequence encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1706–1721, Online. Association for Computational Linguistics.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 35–45, Copenhagen, Denmark. Association for Computational Linguistics.

A cRocoDiLe

CRocoDiLe (Huguet Cabot and Navigli, 2021), based on Elsahar et al. (2018), extracts relational information in Wikipedia abstracts, i.e., the text before the Table of contents. It links the entities present in the text as hyperlinks, together with dates and values, to Wikidata entities using `wikimapper`[4]. The original implementation is compatible with Wikipedia dumps in any language; however, its dates and numbers linker were English-specific. We use regex to extract dates and values in all the languages this work covers.

In the original work, they filtered triplets using NLI. For each triplet, they input the text containing both entities from the Wikipedia abstract, and the triplet in their surface forms, `subject` + `relation` + `object`, separated by the `<sep>` token. If the score was less than 0.75 for the entailment class, it was removed to ensure higher precision. In our work, we set a lower threshold, 0.1, since we further filter triplets using manual annotation or our Critic model.

B Annotation

We employ Mechanical Turk for annotation purposes. Each annotator was paid 0.1$ for every ten instances annotated, constituting 1 HIT, an average of $10 hourly rate. We restrict annotators to countries where each of the languages is spoken, plus the USA. We manually screen annotators in each language separately by having them annotate a small sample of fewer than 10 HITs, and allowing only those who correctly performed the task to annotate the final corpus.

Annotators were presented descriptions for each relation, which they could check at any time by hovering the label or opening the instructions. The English descriptions are:

- **located in the administrative territorial entity**: the item is located on the territory of the following administrative entity
- **country:** sovereign state of this item (not to be used for human beings)
- **instance of:** that class of which this subject is a particular example and member
- **shares border with:** countries or administrative subdivisions, of equal level, that this item borders, either by land or water

[4] https://pypi.org/project/wikimapper/

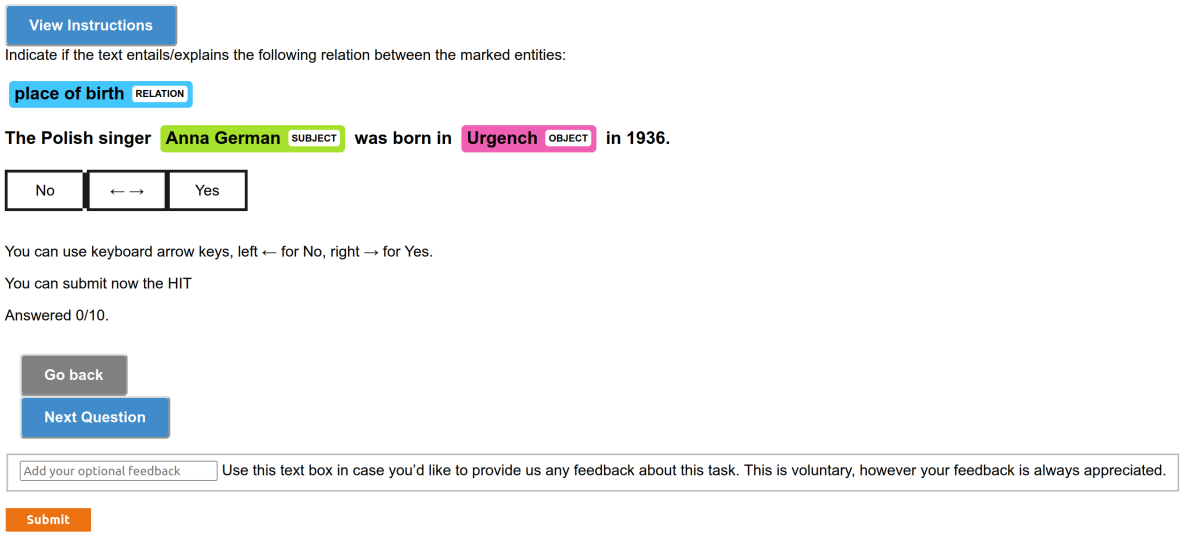

Figure 3: Annotation example from the MT platform. The interface was translated to each language by native speakers.

- **part of:** object of which the subject is a part
- **capital:** seat of government of a country, province, state or other type of administrative territorial entity
- **follows:** immediately prior item in a series of which the subject is a part
- **headquarters location:** city, where an organization's headquarters is or has been situated
- **located in or next to body of water:** sea, lake, river or stream
- **sport:** sport that the subject participates or participated in or is associated with
- **subsidiary:** subsidiary of a company or organization; generally a fully owned separate corporation
- **member of:** organization, club or musical group to which the subject belongs
- **owned by:** owner of the subject
- **manufacturer:** manufacturer or producer of this product
- **genre:** creative work's genre or an artist's field of work (P101)
- **located on terrain feature:** located on the specified landform
- **child:** subject has object as child
- **author:** main creator(s) of a written work (use on works, not humans); use P2093 when Wikidata item is unknown or does not exist
- **named after:** entity or event that inspired the subject's name, or namesake (in at least one language)
- **country of origin:** country of origin of this item (creative work, food, phrase, product, etc.)
- **replaces:** person, state or item replaced
- **inception:** date or point in time when the subject came into existence as defined
- **cast member:** actor in the subject production
- **subclass of:** next higher class or type; all instances of these items are instances of those items; this item is a class (subset) of that item
- **league:** league in which team or player plays or has played in
- **developer:** organization or person that developed the item
- **location:** location of the object, structure or event
- **occupation:** occupation of a person
- **spouse:** the subject has the object as their spouse (husband, wife, partner, etc.)

- **characters:** characters which appear in this item (like plays, operas, operettas, books, comics, films, TV series, video games)
- **notable work:** notable scientific, artistic or literary work, or other work of significance among subject's works
- **place of birth:** most specific known (e.g. city instead of country, or hospital instead of city) birth location of a person, animal or fictional character
- **mouth of the watercourse:** the body of water to which the watercourse drains
- **country of citizenship:** the object is a country that recognizes the subject as its citizen
- **founded by:** founder or co-founder of this organization, religion or place
- **director:** director(s) of film, TV-series, stageplay, video game or similar
- **sibling:** the subject and the object have the same parents (brother, sister, etc.)
- **participant:** person, group of people or organization (object) that actively takes/took part in an event or process (subject)

Figure 3 shows the annotation interface provided to the annotators.

C Results

In this Section, we provide more results concerning our mREBEL model. Specifically, in Figure 4 we provide a heatmap that shows the scores attained by mREBEL$^{T}_{32}$ (without fine-tuning) on each of the 32 relations covered by RED$^{\text{FM}}$, and for each of its 7 languages. Similarly, in Figure 5, we report the scores obtained by the fine-tuned version of mREBEL$^{T}_{32}$. By looking at these two heatmaps, it is easy to identify our model's strengths and weaknesses across relations and languages. We can see how relations such as *named after* or *shares border with* had low scores, probably due to their lower frequency at evaluation time, where a few errors lead to a low score. On the other hand, domain-specific relations such as *cast member*, *league* or *author* show a strong performance on most languages.

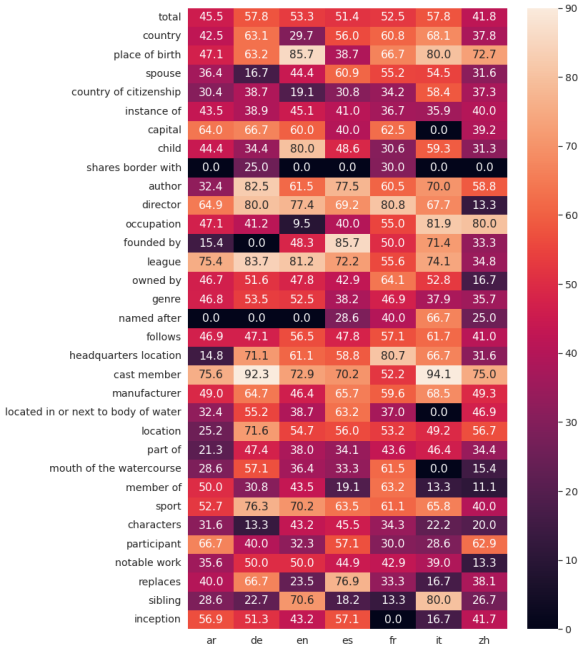

Figure 4: Results for mREBEL$^{T}_{32}$ on RED$^{\text{FM}}$ without fine-tuning.

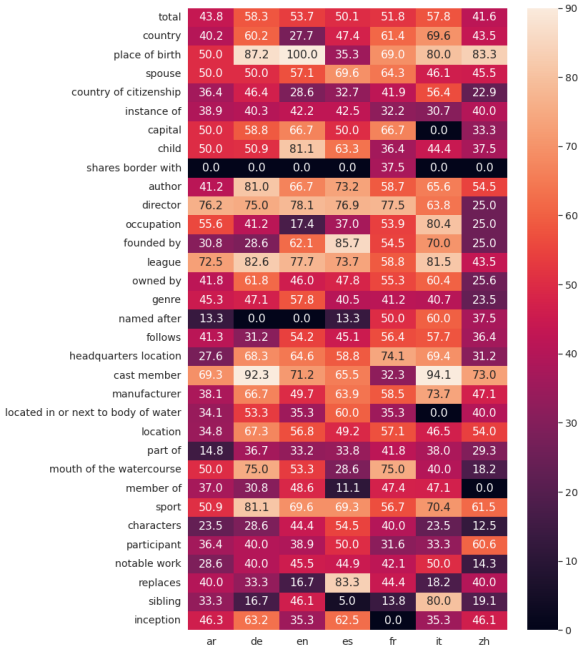

Figure 5: Results for mREBEL$^{T}_{32}$ on RED$^{\text{FM}}$ with fine-tuning.

D Reproducibility

Experiments were performed using a single NVIDIA 3090 GPU with 64GB of RAM and Intel® Core™ i9-10900KF CPU.

The hyperparameters were manually tuned on the validation sets for each dataset, but mostly left at default values for mBART. The ones used for the

	Learning Rate	Warm-up	Batch size	Max Steps
mREBEL$_{400}$	10^{-5}	5000 steps	32	1.6M
mREBEL$_{32}$	10^{-5}	5000 steps	32	+ 264K
mREBEL$_{B400}$	5×10^{-5}	5000 steps	32	1M
SMILER	5×10^{-5}	3000 steps	32	+ 10K
RED$^{\text{FM}}$	10^{-5}	1000 steps	32	+ 10K

Table 9: Hyperparameters for the different datasets. Top half shows used the values used in the pretraining phase, while the bottom part shows those used during fine-tuning.

final results can be found in Table 9.

E Data

In Table 10, we provide data statistics for our RED$^{\text{FM}}$ dataset. Specifically, for each of the 7 languages, we report the number of instances for each relation in the corresponding training, validation and test sets.

	Train						Validation								Test							
	de	**en**	**es**	**fr**	**it**	**Total**	**ar**	**de**	**en**	**es**	**fr**	**it**	**zh**	**Total**	**ar**	**de**	**en**	**es**	**fr**	**it**	**zh**	**Total**
location	896	1071	542	602	528	3639	57	53	99	57	54	55	77	452	73	55	76	61	79	62	100	506
instance of	639	1025	511	603	589	3367	109	74	116	57	78	105	48	587	80	66	116	64	86	115	41	568
country	865	535	646	703	548	3297	136	105	70	79	119	166	104	779	105	127	42	104	142	167	92	779
part of	301	668	353	393	248	1963	35	29	61	39	57	34	42	297	23	23	88	31	29	30	39	263
follows	166	301	210	234	294	1205	52	34	47	27	58	48	35	301	36	16	47	37	35	54	21	246
manufacturer	258	307	218	201	207	1191	92	73	59	42	72	49	61	448	51	51	77	35	74	49	30	367
sport	203	288	172	268	249	1180	45	51	101	41	75	103	11	427	43	39	131	39	61	100	6	419
owned by	199	398	170	198	159	1124	38	28	55	24	33	31	30	239	33	30	49	22	37	28	21	220
located in or next to body of water	169	326	252	159	100	1006	17	4	24	7	10	13	23	98	23	15	16	10	12	2	26	104
member of	125	347	163	159	129	923	19	11	20	16	13	7	19	105	15	20	49	18	23	12	11	148
genre	183	272	145	187	128	915	23	33	59	29	34	72	14	264	28	34	67	44	45	68	11	297
author	150	209	120	139	127	745	29	27	51	23	67	48	28	273	35	58	46	41	44	29	14	267
headquarters location	146	226	110	137	98	717	20	28	28	11	13	17	15	132	9	19	35	9	26	21	17	136
capital	112	191	105	109	104	621	7	9	7	4	9	2	12	50	12	8	4	2	6	3	14	49
child	107	147	104	134	102	594	23	12	16	9	14	12	25	111	15	23	15	28	36	15	33	165
notable work	86	204	61	137	97	585	21	14	47	10	37	34	19	182	19	20	38	17	31	18	4	147
mouth of the watercourse	98	156	205	47	72	578	2	1	15	4	2	6	3	33	4	4	6	3	6	2	3	28
inception	65	335	90	13	69	572	55	10	46	10	3	5	21	150	48	21	36	9	0	9	13	136
named after	139	192	81	80	66	558	10	11	6	4	5	5	7	48	8	6	3	9	10	6	9	51
occupation	172	40	105	139	70	526	34	12	12	13	49	89	12	221	17	18	11	14	43	87	4	194
shares border with	77	56	153	117	63	466	3	2	3	1	4	1	3	17	6	4	4	1	7	1	5	28
participant	48	192	86	66	51	443	6	0	16	5	8	17	6	58	7	2	20	4	13	4	14	64
country of citizenship	111	90	105	71	64	441	23	19	13	18	20	111	29	233	19	26	7	21	18	108	44	243
founded by	81	134	58	97	68	438	16	13	16	2	10	6	7	70	9	4	17	7	10	8	3	58
place of birth	128	94	90	101	23	436	14	18	7	12	12	1	4	68	8	20	3	10	11	2	6	60
replaces	78	118	55	70	70	391	5	4	12	1	6	7	16	51	3	9	11	7	6	7	11	54
cast member	53	145	95	12	43	348	35	38	73	20	20	18	4	208	60	24	102	26	12	16	29	269
league	81	125	42	38	47	333	24	24	33	18	8	24	20	151	35	22	49	18	8	11	12	155
characters	39	113	38	64	46	300	15	9	19	15	16	9	4	87	8	11	24	10	13	13	9	88
director	66	69	37	66	58	296	11	22	16	6	35	26	4	120	20	19	30	12	22	32	7	142
spouse	32	76	49	57	41	255	3	8	8	6	13	7	6	51	8	5	9	11	17	5	9	64
sibling	36	54	23	51	39	203	3	1	5	1	2	1	12	25	4	12	7	9	13	2	5	52
total	5909	8504	5194	5452	4597	29656	982	777	1160	611	956	1129	721	6336	864	811	1235	733	975	1086	663	6367

Table 10: Breakdown for RED^{FM}.