MultiTACRED: A Multilingual Version of the TAC Relation Extraction Dataset

Leonhard Hennig **Philippe Thomas** **Sebastian Möller**
German Research Center for Artificial Intelligence (DFKI)
Speech and Language Technology Lab
{*leonhard.hennig, philippe.thomas, sebastian.moeller*}*@dfki.de*

Abstract

Relation extraction (RE) is a fundamental task in information extraction, whose extension to multilingual settings has been hindered by the lack of supervised resources comparable in size to large English datasets such as *TACRED* (Zhang et al., 2017). To address this gap, we introduce the *MultiTACRED* dataset, covering 12 typologically diverse languages from 9 language families, which is created by machine-translating *TACRED* instances and automatically projecting their entity annotations. We analyze translation and annotation projection quality, identify error categories, and experimentally evaluate fine-tuned pretrained mono- and multilingual language models in common transfer learning scenarios. Our analyses show that machine translation is a viable strategy to transfer RE instances, with native speakers judging more than 83% of the translated instances to be linguistically and semantically acceptable. We find monolingual RE model performance to be comparable to the English original for many of the target languages, and that multilingual models trained on a combination of English and target language data can outperform their monolingual counterparts. However, we also observe a variety of translation and annotation projection errors, both due to the MT systems and linguistic features of the target languages, such as pronoun-dropping, compounding and inflection, that degrade dataset quality and RE model performance.

1 Introduction

Relation extraction (RE), defined as the task of identifying and classifying semantic relationships between entities from text (cf. Figure 1), is a fundamental task in information extraction (Doddington et al., 2004). Extending RE to multilingual settings has recently received increased interest (Zou et al., 2018; Nag et al., 2021; Chen et al., 2022c), both to address the urgent need for more inclusive NLP systems that cover more languages than just English (Ruder et al., 2019; Hu et al., 2020), as well as to investigate language-specific phenomena and challenges relevant to this task. The main bottleneck for multilingual RE is the lack of supervised resources, comparable in size to large English datasets (Riedel et al., 2010; Zhang et al., 2017), as annotation for new languages is very costly. Most of the few existing multilingual RE datasets are distantly supervised (Köksal and Özgür, 2020; Seganti et al., 2021; Bhartiya et al., 2022), and hence suffer from noisy labels that may reduce the prediction quality of models (Riedel et al., 2010; Xie et al., 2021). Available fully-supervised datasets are small, and cover either very few domain-specific relation types (Arviv et al., 2021; Khaldi et al., 2022), or only a small set of languages (Nag et al., 2021).

To address this gap, and to incentivize research on supervised multilingual RE, we introduce a multilingual version of one of the most prominent supervised RE datasets, *TACRED* (Zhang et al., 2017). *MultiTACRED* is created by machine-translating *TACRED* instances and automatically projecting their entity annotations. Machine translation is a popular approach for generating data in cross-lingual learning (Hu et al., 2020; Nag et al., 2021). Although the quality of machine-translated data may be lower due to translation and alignment errors (Yarmohammadi et al., 2021), it has been shown to be beneficial for classification and structured prediction tasks (Hu et al., 2020; Ozaki et al., 2021; Yarmohammadi et al., 2021).

The *MultiTACRED* dataset we present in this work covers 12 languages from 9 language families.[1] We select typologically diverse languages which span a large set of linguistic phenomena such as compounding, inflection and pronoun-drop,

[1] *MultiTACRED* includes the following language families / languages: German (Germanic); Finnish, Hungarian (Uralic); Spanish, French (Romance); Arabic (Semitic); Hindi (Indo-Iranic); Japanese (Japonic); Polish, Russian (Slavic); Turkish (Turkic); Chinese (Sino-Tibetan).

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 3785–3801
July 9-14, 2023 ©2023 Association for Computational Linguistics

and for which a monolingual pretrained language model is available. We automatically and manually analyze translation and annotation projection quality in all target languages, both in general terms and with respect to the RE task, and identify typical error categories for alignment and translation that may affect model performance. We find that overall translation quality is judged to be quite good with respect to the RE task, but that e.g. pronoun-dropping, coordination and compounding may cause alignment and semantic errors that result in erroneous instances. In addition, we experimentally evaluate fine-tuned pretrained mono- and multilingual language models (PLM) in common training scenarios, using source language (English), target language, or a mixture of both as training data. We also evaluate an English data fine-tuned model on back-translated test instances to estimate the effect of noise introduced by the MT system on model performance. Our results show that in-language training works well, given a suitable PLM. Cross-lingual zero-shot transfer is acceptable for languages well-represented in the multilingual PLM, and combining English and target language data for training considerably improves performance across the board.

To summarize, our work aims to answer the following research questions: Can we reaffirm the usefulness of MT and cross-lingual annotation projection, in our study for creating large-scale, high quality multilingual datasets for RE? How do pretrained mono- and multilingual encoders compare to each other, in within-language as well as cross-lingual evaluation scenarios? Answers to these questions can provide insights for understanding language-specific challenges in RE, and further research in cross-lingual representation and transfer learning. The contributions of this paper are:

- We introduce *MultiTACRED*, a translation of the widely used, large-scale *TACRED* dataset into 12 typologically diverse target languages: Arabic, German, Spanish, French, Finnish, Hindi, Hungarian, Japanese, Polish, Russian, Turkish, and Chinese.

- We present an evaluation of monolingual, cross-lingual, and multilingual models to evaluate target language performance for all 12 languages.

- We present insights into the quality of machine translation for RE, analyzing alignment as well as language-specific errors.

2 Translating TACRED

We first briefly introduce the original *TACRED* dataset, and then describe the language selection and automatic translation process. We wrap up with a description of the analyses we conduct to verify the translation quality.

2.1 The TACRED dataset

The *TAC Relation Extraction Dataset*[2], introduced by Zhang et al. (2017), is a fully supervised dataset of sentence-level binary relation mentions. It consists of 106k sentences with entity mention pairs collected from the TAC KBP[3] evaluations 2009–2014, with the years 2009 to 2012 used for training, 2013 for development, and 2014 for testing. Each sentence is annotated with a head and a tail entity mention, and labeled with one of 41 person- and organization-oriented relation types, e.g. *per:title*, *org:founded*, or the label *no_relation* for negative instances. About 79.5% of the examples are labeled as *no_relation*.[4] All relation labels were obtained by crowdsourcing, using Amazon Mechanical Turk. Recent work by Alt et al. (2020) and Stoica et al. (2021) improved upon the label quality of the crowd annotations by re-annotating large parts of the dataset.

2.2 Automatic Translation

We translate the complete *train*, *dev* and *test* splits of *TACRED* into the target languages, and in addition back-translate the *test* split into English to generate machine-translated English test data. Each instance in the original *TACRED* dataset is a list of tokens, with the head and tail entity arguments of the potential relation specified via token offsets. For translation, we concatenate tokens with whitespace and convert head and tail entity offsets into XML-style markers to denote the arguments' boundaries, as shown in Figure 1. We use the commercial services of DeepL[5] and Google[6], since both offer the functionality to preserve XML tag markup. Since API costs are similar, we use DeepL for most languages, and only switch to Google for

[2] https://catalog.ldc.upenn.edu/LDC2018T24, under a LDC license

[3] https://tac.nist.gov/2017/KBP/index.html

[4] The first row of Table 6 in Appendix A summarizes key statistics of the dataset.

[5] https://api.deepl.com/v2/translate

[6] https://translation.googleapis.com/language/translate/v3

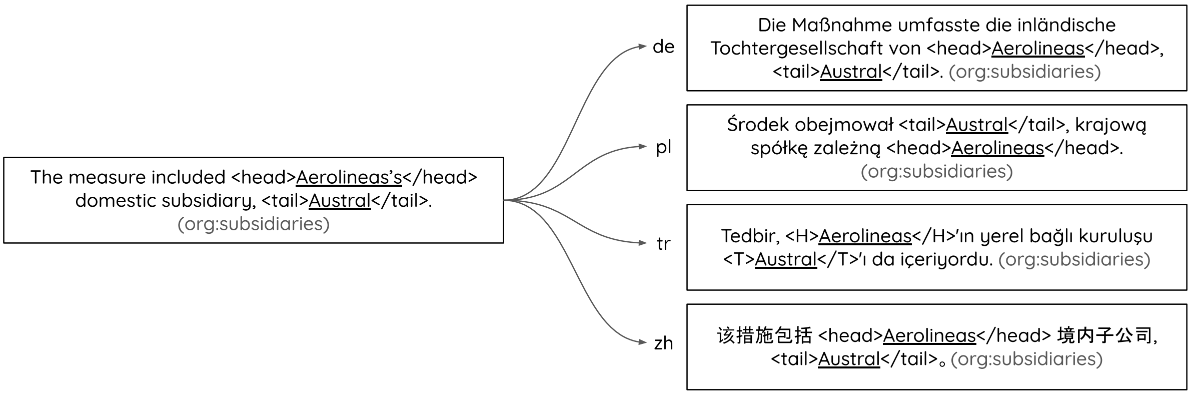

The measure included <head>Aerolineas's</head> domestic subsidiary, <tail>Austral</tail>. (org:subsidiaries)
de
Die Maßnahme umfasste die inländische Tochtergesellschaft von <head>Aerolineas</head>, <tail>Austral</tail>. (org:subsidiaries)
pl
Środek obejmował <tail>Austral</tail>, krajową spółkę zależną <head>Aerolineas</head>. (org:subsidiaries)
tr
Tedbir, <H>Aerolineas</H>'ın yerel bağlı kuruluşu <T>Austral</T>'ı da içeriyordu. (org:subsidiaries)
zh
该措施包括 <head>Aerolineas</head> 境内子公司, <tail>Austral</tail>。(org:subsidiaries)

Figure 1: Example translations from English to German, Polish, Turkish and Chinese with XML markup for the head and tail entities to project relation argument annotations.

languages not supported by DeepL (at the time we were running the MT). We validate the translated text by checking the syntactic correctness of the XML tag markup, and discard translations with invalid tag structure, e.g. missing or invalid head or tail tag pairs.

After translation, we tokenize the translated text using language-specific tokenizers.[7] Finally, we store the translated instances in same JSON format as the original *TACRED* English dataset, with fields for tokens, entity types and offsets, label and instance id. We can then easily apply the label corrections provided by e.g. Alt et al. (2020) or Stoica et al. (2021) to any target language dataset by applying the respective patch files.

We select target languages to cover a wide set of interesting linguistic phenomena, such as compounding (e.g., German), inflection/derivation (e.g., German, Turkish, Russian), pronoun-dropping (e.g., Spanish, Finnish, Polish), and varying degrees of synthesis (e.g., Turkish, Hungarian vs. Chinese). We also try to ensure that there is a monolingual pretrained language model available for each language, which is the case for all languages except Hungarian. The final set of languages in *MultiTACRED* is: German, Finnish, Hungarian, French, Spanish, Arabic, Hindi, Japanese, Chinese, Polish, Russian, and Turkish. Table 6 in Appendix A lists key statistics per language.

2.3 Translation Quality Analysis

To verify the overall quality of the machine-translated data, we also manually inspect translations. For each language, we randomly sample 100 instances from the *train* split. For each sample instance, we display the source (English) text with entity markup (see Figure 1 for the format), the target language text with entity markup, and the relation label.

We then ask native speakers to judge the translations by answering two questions: (Q1) Does the translated text meaningfully preserve the semantic relation of the English original, regardless of minor translation errors?[8] (Q2) Is the overall translation linguistically acceptable for a native speaker? Human judges are instructed to read both the English source and the translation carefully, and then to answer the two questions with either *yes* or *no*. They may also add free-text comments, e.g. to explain their judgements or to describe translation errors. The samples of each language are judged by a single native speaker. Appendix B gives additional details.

In addition, we conduct a manual analysis of the automatically discarded translations, using a similar-sized random sample from the German, Russian and Turkish *train* splits, to identify possible reasons and error categories. These analyses are performed by a single trained linguist per language, who is also a native speaker of that language, with joint discussions to synthesize observations. Results of both analyses are presented in Section 4.1.

3 Experiments

In this section, we describe the experiments we conduct to answer the research questions "How does the performance of language-specific models compare to the English original?" and "How does

[7] See Appendix A for details.

[8] If necessary, human judges are first introduced to the task of relation extraction. They are also given the list of relations and their official definitions for reference.

the performance of language-specific models compare to multilingual models such as mBERT trained on the English source data? How does the performance change when including target-language data for training". We first introduce the training scenarios, and then give details on choice of models and hyperparameters, as well as the training process.

3.1 Training scenarios

We evaluate the usefulness of the translated datasets by following the most prevalent approach of framing RE as a sentence-level supervised multi-class classification task. Formally, given a relation set $\mathcal{R}$ and a text $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$ (where $x_1, \cdots, x_n$ are tokens) with two disjoint spans $\boldsymbol{e}_h = [x_i, \ldots, x_j]$ and $\boldsymbol{e}_t = [x_k, \ldots, x_l]$ denoting the head and tail entity mentions, RE aims to predict the relation $r \in \mathcal{R}$ between $\boldsymbol{e}_h$ and $\boldsymbol{e}_t$, or assign the *no_relation* class if no relation in $\mathcal{R}$ holds. Similar to prior work (e.g., Nag et al. (2021)), we evaluate relation extraction models in several different transfer learning setups, which are described next.

Monolingual We evaluate the performance of language-specific PLMs for each of the 12 target languages, plus English, where the PLM is supervisedly fine-tuned on the *train* split of the respective language.

Cross-lingual We evaluate the performance of a multilingual mBERT model on the test split of each of the 12 target languages, plus English, after training on the English *train* split.

Mixed / Multilingual We evaluate the performance of a multilingual mBERT model on the test split of each of the 12 target languages, after training on the complete English *train* split and a variable portion of the *train* split of the target language, as suggested e.g. by Nag et al. (2021). We vary the amount of target language data in {5%,10%,20%,30%,40%,50%,100%} of the available training data. When using 100%, we are effectively doubling the size of the training set, and "duplicating" each training instance.

Back-translation Finally, we also evaluate the performance of a BERT model fine-tuned on the original (untranslated) English train split on the test sets obtained by back-translating from each target language.

3.2 Training Details and Hyperparameters

We implement our experiments using the Hugging Face (HF) Transformers library (Wolf et al., 2020), Hydra (Yadan, 2019) and PyTorch (Paszke et al., 2019).[9] Due to the availability of pretrained models for many languages and to keep things simple, we use BERT as the base PLM (Devlin et al., 2019).

We follow Baldini Soares et al. (2019) and enclose the subject and object entity mentions with special token pairs, modifying the input to become "`[HEAD_START] subject [HEAD_END] ...[TAIL_START] object [TAIL_END]`". In addition, we append the entity types of subject and object to the input text as special tokens, after a separator token: "`...[SEP] [HEAD=type] [SEP] [TAIL=type]`", where `type` is the entity type of the respective argument. We use the final hidden state representation of the `[CLS]` token as the fixed length representation of the input sequence that is fed into the classification layer.

We train with batch size of 8 for 5 epochs, and optimize for cross-entropy. The maximum sequence length is 128 for all models. We use AdamW with a scenario-specific learning rate, no warmup, β_1 = 0.9, $\beta_2 = 0.999$, $\epsilon = 1e-8$, and linear decay of the learning rate. Other hyperparameter values, as well as scenario-specific learning rates and HF model identifiers for the pretrained BERT models, are listed in Appendix C.

We use micro-F1 as the evaluation metric, and report the median result of 5 runs with different, fixed random seeds. For all experiments, we use the revised version of *TACRED* presented by Alt et al. (2020), which fixes a large portion of the *dev* and *test* labels.[10] We report scores on the test set in the respective target language, denoted as $test_L$. Due to the automatic translation and validation, training and test sets differ slightly across languages, and absolute scores are thus not directly comparable across languages. We therefore also report scores on the intersection test set of instances available in all languages ($test_\cap$). This test set contains 11,874 instances, i.e. 76.6% of the original test set (see also Table 6).

4 Results and Discussion

We first present some insights into translation quality, and then discuss the performance of models for the different training scenarios.

[9] We make our code publicly available at `https://github.com/DFKI-NLP/MultiTACRED` for better reproducibility.

[10] Since both Alt et al. (2020) and (Stoica et al., 2021) provide fixes as patch files to the original dataset, it is trivial to repeat our experiments using the original or the *Re-TACRED* version of the data.

4.1 Translation Quality

Automatic validation As described in Section 2.2, we validate the target language translation by checking whether the entity mention tag markup was correctly transferred. On average, 2.3% of the instances were considered invalid after translation. By far the largest numbers of such errors occurred when translating to Japanese (9.6% of translated instances), followed by Chinese (4.5%) and Spanish (3.8%). Table 6 in Appendix A gives more details, and shows the number of valid translations for each language, per split and also for the back-translation of the test split. Back-translation incurred only half as many additional errors as compared to the initial translation of the test split into the target language, presumably due to the fact that 'hard' examples had already been filtered out during the first translation step.

The validation basically detects two types of alignment errors - missing and additional alignments. An alignment may be missing in the case of pro-drop languages, where the argument is not realized in the translation (e.g. Spanish, Chinese), or in compound noun constructions in translations (e.g. in German). In other cases, the aligner produces multiple, disjoint spans for one of the arguments, e.g. in the case of coordinated conjunctions or compound constructions with different word order in the target language (e.g. in Spanish, French, Russian). Table 8 in Appendix D lists more examples for the most frequent error categories we observed.

Manual Validation Table 1 shows the results of the manual analysis of translations. With regards to Q1, on average 87.5% of the translations are considered to meaningfully express the relation, i.e. as in the original text. Overall translation quality is judged to be good for 83.7% of the sampled instances on average across languages. The most frequent error types noted by the annotators are again alignment errors, such as aligning a random (neighboring) token from the sentence with an English pronoun argument in pronoun-dropping languages (e.g. Polish, Chinese), and non-matching spans (inclusion/exclusion of tokens in the aligned span). Similar errors have also been observed in a recent study by Chen et al. (2022b). In highly inflecting languages such as Finnish or Turkish, the aligned entity often changes morphologically (e.g. possessive/case suffixes).[11] Other typical errors are uncommon/wrong word choices, (e.g. due to missing or wrongly interpreted sentence context), and the omission of parts of the original sentence. Less frequent errors include atypical input which was not translated correctly (e.g. sentences consisting of a list of sports results), and non-English source text (approx. 1% of the data, see also Stoica et al. (2021)). Table 8 also lists examples for these error categories.

Language	Q1 (yes)	Q2 (yes)
ar	85%	92%
de	100%	91%
es	78%	91%
fi	82%	81%
fr	92%	93%
hi	89%	67%
hu	89%	48%
ja	74%	89%
pl	73%	93%
ru	98%	89%
tr	99%	90%
zh	91%	80%
Avg	87.5%	83.7%

Table 1: Translation quality, as judged by native speakers. (Q1) Does the translated text meaningfully express the semantic relation of the English original, regardless of minor translation errors? (Q2) Is the overall translation linguistically acceptable for a native speaker?

4.2 Model Performance

Monolingual Table 2 shows the results for the monolingual setting. The English BERT model achieves a reference median micro-F1 score of 77.1, which is in line with similar results for fine-tuned PLMs. (Alt et al., 2020; Chen et al., 2022a; Zhou and Chen, 2022) Micro-F1 scores for the other languages range from 71.8 (Hungarian) to 76.4 (Finnish), with the notable exception of Hindi, where the fine-tuned BERT model only achieves a micro-F1 score of 65.1[12]. As discussed in Section 3.2, results are not directly comparable across languages. However, the results in Table 2 show that language-specific models perform reasonably well for many of the evaluated languages.[13] Their

[11] Inflection and compounding both ideally could be solved by introducing alignment/argument span boundaries at the morpheme level, but this in turn may raise issues with e.g. PLM tokenization and entity masking.

[12] See also Appendix C for an additional discussion of Hindi performance issues

[13] However, as various researchers have pointed out, model performance may be over-estimated, since the models may be

Test set	en	ar	de	es	fi	fr	hi	hu	ja	pl	ru	tr	zh
$test_L$	77.1	74.2	74.1	75.7	76.4	75.0	65.1	71.8	71.8	73.7	73.7	74.2	75.4
$test_\cap$	77.5	74.5	74.6	76.1	76.6	75.4	65.9	72.4	72.5	74.3	74.8	74.5	75.3

Table 2: Micro-F1 scores on the TACREV dataset for the monolingual setting. The table shows the median micro-F1 score across 5 runs, on the *test* split of the target language ($test_L$), and on the intersection of test instances available in all languages ($test_\cap$).

lower performance may be due to several reasons: translation errors, smaller train and test splits because of the automatic validation step, the quality of the pre-trained BERT model, as well as language-specific model errors.

Results on the intersection test set $test_\cap$ are slightly higher on average, as compared to $test_L$. Relative differences to English, and the overall 'ranking' of language-specific results, remain approximately the same. This reaffirms the performance differences between languages observed on $test_L$. It also suggests that the intersection test set contains fewer challenging instances. For Hindi, these results, in combination with the low manual evaluation score of 67% correct translations, suggest that the translation quality is the main reason for the performance loss.

We conclude that for the monolingual scenario, machine translation is a viable strategy to generate supervised data for relation extraction for most of the evaluated languages. Fine-tuning a language-specific PLM on the translated data yields reasonable results that are not much lower than those of the English model for many tested languages.

Cross-lingual In the cross-lingual setting, micro-F1 scores are lower than in the monolingual setting for many languages (see Table 3). The micro-F1 scores for languages well-represented in mBERT's pretraining data (e.g., English, German, Chinese) are close to their monolingual counterparts, whereas for languages like Arabic, Hungarian, Japanese, or Turkish, we observe a loss of 4.7 to 9.7 F1 points. This is mainly due to a much lower recall, for example, the median recall for Japanese is only 51.3. The micro-F1 scores are highly correlated with the pretraining data size of each language in mBERT: The Spearman rank correlation coefficient of micro-F1 L_T scores with the WikiSize reported in Wu and Dredze (2020) is $r_s = 0.82$, the Pearson correlation coefficient is $r_p = 0.78$. Hence, languages which are less well represented in mBERT's pretraining data exhibit worse relation extraction performance, as they don't benefit as much from the pretraining.

Precision, Recall and F1 on the intersection test set $test_\cap$ are again slightly better on average than the scores on $test_L$. For Hindi, our results reaffirm the observations made by Nag et al. (2021) for cross-lingual training using only English training data. Our results for RE also confirm prior work on the effectiveness of cross-lingual transfer learning for other tasks (e.g., Conneau et al. (2020); Hu et al. (2020). While results are lower than in the monolingual setting, they are still very reasonable for well-resourced languages such as German or Spanish, with the benefit of incurring no translation at all for training. However, for languages that are less well-represented in mBERT, using a language-specific PLM in combination with in-language training data produces far better results.

Mixed/Multilingual Table 4 shows the results obtained when training on both English and varying amounts of target language data. We can observe a considerable increase of mBERT's performance for languages that are not well represented in mBERT's pretraining data, such as e.g. Hungarian. These languages benefit especially from adding in-language training data, in some cases even surpassing the performance of their respective monolingual model. For example, mBERT trained on the union of the English and the complete Japanese *train* splits achieves a micro-F1 score of 73.3, 11.2 points better than the cross-lingual score of 62.1 and 1.5 points better than the 71.8 obtained by the monolingual model on the same test data. Languages like German, Spanish, and French don't really benefit from adding small amounts of in-language training data in our evaluation, but show some improvements when adding 100% of the target language training data (last row), i.e. when essentially doubling the size of the training data. Other languages, like Finnish or Turkish, show improvements over the cross-lingual baseline, but don't reach the performance of their monolingual counterpart.

affected by "translationese" (Riley et al., 2020; Graham et al., 2020).

Test set / Wikisize	Metric	en	ar	de	es	fi	fr	hi	hu	ja	pl	ru	tr	zh
$test_L$	P	76.7	72.1	75.2	74.0	76.7	74.3	76.1	76.5	78.6	76.9	70.6	73.6	73.2
	R	77.5	60.3	74.0	73.9	64.9	73.9	53.0	59.7	51.3	70.0	74.6	57.4	70.0
	F1	77.1	65.7	74.6	73.9	70.3	74.1	62.5	67.1	62.1	73.3	72.6	64.5	71.6
$test_\cap$	P	76.5	73.2	75.5	74.8	78.3	75.0	76.5	76.6	79.2	77.1	70.6	73.4	73.8
	R	78.3	61.6	74.3	75.3	65.1	74.3	54.3	60.7	50.8	71.1	75.3	58.1	69.9
	F1	77.4	66.9	74.9	75.0	71.1	74.6	63.5	67.7	61.9	74.0	72.9	64.9	71.8
WikiSize	$\log_2$(MB)	14	10	12	12	9	12	7	10	11	11	12	9	11

Table 3: Micro-Precision, Recall and F1 scores on the TACREV dataset for the cross-lingual setting. The table shows the median scores across 5 runs, on the translated *test* split of the target language ($test_L$) and on the intersection of test instances available in all languages ($test_\cap$), when training mBERT on the English *train* split. For reference, the table also shows the size of mBERT's training data in a given language (Wikisize, as log_2(MegaBytes), taken from Wu and Dredze (2020)). Languages with less pretraining data in mBERT suffer a larger performance loss.

In-lang data (%)	ar	de	es	fi	fr	hi	hu	ja	pl	ru	tr	zh	$\overline{\Delta}$
-	65.7	74.6	73.9	70.3	74.1	62.5	67.1	62.1	73.3	72.6	64.5	71.6	-
5	68.7	74.9	73.9	70.5	74.3	67.0	68.8	69.2	72.2	73.0	67.8	72.5	+1.7
10	69.0	74.5	73.7	70.4	73.6	68.0	68.9	70.6	72.0	72.7	68.9	73.0	+1.9
20	71.0	74.4	74.5	72.2	73.9	69.9	70.2	71.7	73.3	73.3	69.2	73.0	+2.9
30	71.4	74.8	74.8	72.3	74.2	70.1	71.0	72.3	72.9	73.2	70.1	73.7	+3.2
40	71.2	74.3	74.5	72.1	73.9	70.4	70.8	71.6	73.0	73.1	70.3	74.0	+3.1
50	71.2	74.7	74.4	73.0	74.4	71.8	70.9	72.6	73.1	73.3	70.3	74.8	+3.5
100	73.5	75.8	75.9	73.5	75.6	72.4	72.4	73.3	74.3	75.6	71.6	75.4	+4.7

Table 4: Micro-F1 scores on the TACREV dataset for the mixed/multilingual setting. The table shows the median micro-F1 score across 5 runs, on the translated *test* split of the target language, when training mBERT on the full English *train* split and various portions, from 5% to 100%, of the translated target language *train* split. The last column shows the mean improvement across languages, compared to the cross-lingual baseline. Micro-F1 scores improve when adding in-language training data for languages not well represented in mBERT, while other languages mainly benefit when using all of the English and in-language data, i.e. essentially doubling the amount of training data (last row).

Our results confirm observations made by Nag et al. (2021), who also find improvements when training on a mixture of gold source language data and projected silver target language data. For the related task of event extraction, Yarmohammadi et al. (2021) also observe that the combination of data projection via machine translation and multilingual PLMs can lead to better performance than any one cross-lingual strategy on its own.

Back-translation Finally, Table 5 shows the performance of the English model evaluated on the back-translated test splits of all target languages. Micro-F1 scores range from 69.6 to 76.1, and are somewhat lower than the score of 77.1 achieved by the same model on the original test set. For languages like German, Spanish, and French, scores are very close to the original, while for Arabic and Hungarian, we observe a loss of approximately 7 percentage points. These differences may be due to the different quality of the MT systems per language pair, but can also indicate that the model cannot always handle the linguistic variance introduced by the back-translation.

5 Related Work

Multilingual RE Datasets Prior work has primarily focused on the creation of distantly supervised datasets. Dis-Rex (Bhartiya et al., 2022) and RelX-Distant (Köksal and Özgür, 2020) are large, Wikipedia-based datasets, but cover only 4 resp. 5 European languages. SMiLER (Seganti et al., 2021) covers 14 European languages, but is very imbalanced, both in terms of relation coverage in the different languages and training data per language (Chen et al., 2022c).

Manually supervised datasets include BizRel (Khaldi et al., 2022), consisting of 25.5K sentences labeled with 5 business-oriented relation types, in French, English, Spanish and Chinese, and the IndoRE dataset of 32.6K sentences covering 51 Wikidata relations, in Bengali,

Language	ar	de	es	fi	fr	hi	hu	ja	pl	ru	tr	zh
F1	69.6	76.1	75.8	73.6	75.9	73.3	70.0	72.2	74.7	74.0	72.1	74.8

Table 5: Median micro-F1 scores across 5 runs of the English BERT model evaluated on the back-translated *test* splits of all languages. Compared to the micro-F1 score of 77.1 on the untranslated English test set, back-translation results are somewhat lower, due to MT system quality and the linguistic variance introduced by the back-translation.

Hindi, Telugu and English (Nag et al., 2021). The IndoRE dataset uses MT to transfer manually labeled examples from English to the three other languages, but implements a heuristic to project entity annotations, without any verification step. Other datasets are very small: The RelX dataset contains a manually translated parallel test set of 502 sentences (Köksal and Özgür, 2020). Arviv et al. (2021) create a small parallel RE dataset of 533 sentences by sampling from *TACRED* and translating into Russian and Korean. For the related task of event extraction, datasets worth mentioning are the multilingual ACE 2005 dataset (Walker et al., 2006), the TAC multilingual event extraction dataset (Ellis et al., 2016), and the work of Yarmohammadi et al. (2021).

Machine Translation for Cross-lingual Learning MT is a popular approach to address the lack of data in cross-lingual learning (Hu et al., 2020; Nag et al., 2021). There are two basic options - translating target language data to a well-resourced source language at inference time and applying a model trained in the source language (Asai et al., 2018; Cui et al., 2019; Hu et al., 2020), or translating source language training data to the target language, while also projecting any annotations required for training, and then training a model in the target language (Khalil et al., 2019; Yarmohammadi et al., 2021; Kolluru et al., 2022). Both approaches depend on the quality of the MT system, with translated data potentially suffering from translation or alignment errors (Aminian et al., 2017; Ozaki et al., 2021; Yarmohammadi et al., 2021). With very few exceptions, using MT for multilingual RE remains underexplored (Faruqui and Kumar, 2015; Zou et al., 2018; Nag et al., 2021).

Multilingual RE Previous work in cross- and multilingual RE has explored a variety of approaches. Kim et al. (2014) proposed cross-lingual annotation projection, while Faruqui and Kumar (2015) machine-translate non-English sentences to English, and then project the relation phrase back to the source language for the task of Open RE. Verga et al. (2016) use multilingual word embeddings to extract relations from Spanish text without using Spanish training data. In a related approach, Ni and Florian (2019) describe an approach for cross-lingual RE that is based on bilingual word embedding mapping. Lin et al. (2017) employ convolutional networks to extract relation embeddings from texts, and propose cross-lingual attention between relation embeddings to model cross-lingual information consistency. Chen et al. (2022c) introduce a prompt-based model, which requires only the translation of prompt verbalizers. Their approach thus is especially useful in few- and zero-shot scenarios.

6 Conclusion

We introduced a multilingual version of the large-scale *TACRED* relation extraction dataset, obtained via machine translation and automatic annotation projection. Baseline experiments with in-language as well as cross-lingual transfer learning models showed that MT is a viable strategy to transfer sentence-level RE instances and span-level entity annotations to typologically diverse target languages, with target language RE performance comparable to the English original for many languages.

However, we observe that a variety of errors may affect the translations and annotation alignments, both due to the MT system and the linguistic features of the target languages (e.g., compounding, high level of synthesis). *MultiTACRED* can thus serve as a starting point for deeper analyses of annotation projection and RE challenges in these languages. For example, we would like to improve our understanding of RE annotation projection for highly inflectional/synthetic languages, where token-level annotations are an inadequate solution. In addition, constructing original-language test sets to measure the effects of translationese remains an open challenge.

We plan to publish the translated dataset for the research community, depending on LDC requirements for the original *TACRED* and the underlying

TAC corpus. We will also make publicly available the code for the automatic translation, annotation projection, and our experiments.

Limitations

A key limitation of this work is the dependence on a machine translation system to get high-quality translations and annotation projections of the dataset. Depending on the availability of language resources and the MT model quality for a given language pair, the translations we use for training and evaluation may be inaccurate, or be affected by translationese, possibly leading to overly optimistic estimates of model performance. In addition, since the annotation projection for relation arguments is completely automatic, any alignment errors of the MT system will yield inaccurate instances. Alignment is at the token-level, rendering it inadequate for e.g. compounding or highly inflectional languages. Due to the significant resource requirements of constructing adequately-sized test sets, another limitation is the lack of evaluation on original-language test instances. While we manually validate and analyze sample translations in each target language (Section 4.1) for an initial exploration of MT effects, these efforts should be extended to larger samples or the complete test sets. Finally, we limited this work to a single dataset, which was constructed with a specific set of target relations (person- and organization-related), from news and web text sources. These text types and the corresponding relation expressions may be well reflected in the training data of current MT systems, and thus easier to translate than relation extraction datasets from other domains (e.g., biomedical), or other text types (e.g., social media). The translated examples also reflect the source language's view of the world, not how the relations would necessarily be formulated in the target language (e.g., use of metaphors, or ignorance of cultural differences).

Ethics Statement

We use the data of the original *TACRED* dataset "as is". Our translations thus reflect any biases of the original dataset and its construction process, as well as biases of the MT models (e.g., rendering gender-neutral English nouns to gendered nouns in a given target language). The authors of the original *TACRED* dataset (Zhang et al., 2017) have not stated measures that prevent collecting sensitive text. Therefore, we do not rule out the possible risk of sensitive content in the data. Furthermore, we utilize various BERT-based PLMs in our experiments, which were pretrained on a wide variety of source data. Our models may have inherited biases from these pretraining corpora.

Training jobs were run on a machine with a single NVIDIA RTX6000 GPU with 24 GB RAM. Running time per training/evaluation is approximately 1.5 hours for the monolingual and cross-lingual models, and up to 2 hours for the mixed/multilingual models that are trained on English and target language data.

Acknowledgements

We would like to thank David Harbecke, Aleksandra Gabryszak, Nils Feldhus and the anonymous reviewers for their valuable comments and feedback on the paper. We are also very grateful to all the helpful annotators who evaluated the translations: Ammer Ayach, Yuxuan Chen, Nicolas Delinte, Aleksandra Gabryszak, Maria Gonzalez Garcia, Elif Kara, Tomohiro Nishiyama, Akseli Reunamo, Kinga Schumacher, Akash Sinha, and Tatjana Zeen. Finally, we'd like to thank Gabriel Kressin and Phuc Tran Truong for their help with the code base and running the translations and experiments. This work has been supported by the German Federal Ministry for Economic Affairs and Climate Action as part of the project PLASS (01MD19003E), and by the German Federal Ministry of Education and Research as part of the projects CORA4NLP (01IW20010) and Text2Tech (01IS22017B).

References

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online. Association for Computational Linguistics.

Maryam Aminian, Mohammad Sadegh Rasooli, and Mona Diab. 2017. Transferring semantic roles using translation and syntactic information. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 13–19, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D. Manning. 2014. Combining distant and partial supervision for relation extraction. In *Proceedings of the 2014 Conference on Empirical Methods in*

Natural Language Processing (EMNLP), pages 1556–1567, Doha, Qatar. Association for Computational Linguistics.

Ofir Arviv, Dmitry Nikolaev, Taelin Karidi, and Omri Abend. 2021. On the relation between syntactic divergence and zero-shot performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4803–4817, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *ArXiv*, abs/1809.03275.

Livio Baldini Soares, Nicholas FitzGerald, Jeffrey Ling, and Tom Kwiatkowski. 2019. Matching the Blanks: Distributional Similarity for Relation Learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy. Association for Computational Linguistics.

Abhyuday Bhartiya, Kartikeya Badola, and Mausam . 2022. DiS-ReX: A multilingual dataset for distantly supervised relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 849–863, Dublin, Ireland. Association for Computational Linguistics.

Xiang Chen, Ningyu Zhang, Ningyu Zhang, Xin Xie, Shumin Deng, Yunzhi Yao, Chuanqi Tan, Fei Huang, Luo Si, and Huajun Chen. 2022a. Knowprompt: Knowledge-aware prompt-tuning with synergistic optimization for relation extraction. *Proceedings of the ACM Web Conference 2022*.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2022b. Frustratingly easy label projection for cross-lingual transfer. *CoRR*, abs/2211.15613.

Yuxuan Chen, David Harbecke, and Leonhard Hennig. 2022c. 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, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2019. Cross-lingual machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1586–1595, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Joe Ellis, Jeremy Getman, Dana Fore, Neil Kuster, Zhiyi Song, Ann Bies, and Stephanie Strassel. 2016. Overview of linguistic resources for the tac kbp 2016 evaluations: Methodologies and results. In *Proceedings of TAC 2016*.

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*, pages 1351–1356, Denver, Colorado. Association for Computational Linguistics.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2020. Statistical power and translationese in machine translation evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 72–81, Online. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Hadjer Khaldi, Farah Benamara, Camille Pradel, Grégoire Sigel, and Nathalie Aussenac-Gilles. 2022. How's business going worldwide ? a multilingual annotated corpus for business relation extraction. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3696–3705, Marseille, France. European Language Resources Association.

Talaat Khalil, Kornel Kiełczewski, Georgios Christos Chouliaras, Amina Keldibek, and Maarten Versteegh.

2019. Cross-lingual intent classification in a low resource industrial setting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6419–6424, Hong Kong, China. Association for Computational Linguistics.

Seokhwan Kim, Minwoo Jeong, Jonghoon Lee, and Gary Geunbae Lee. 2014. Cross-lingual annotation projection for weakly-supervised relation extraction. *ACM Transactions on Asian Language Information Processing*, 13(1).

Abdullatif Köksal and Arzucan Özgür. 2020. The RELX dataset and matching the multilingual blanks for cross-lingual relation classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 340–350, Online. Association for Computational Linguistics.

Keshav Kolluru, Muqeeth Mohammed, Shubham Mittal, Soumen Chakrabarti, and Mausam . 2022. Alignment-augmented consistent translation for multilingual open information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2502–2517, Dublin, Ireland. Association for Computational Linguistics.

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)*, pages 34–43, Vancouver, Canada. Association for Computational Linguistics.

Arijit Nag, Bidisha Samanta, Animesh Mukherjee, Niloy Ganguly, and Soumen Chakrabarti. 2021. A data bootstrapping recipe for low-resource multilingual relation classification. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 575–587, Online. Association for Computational Linguistics.

Jian Ni and Radu Florian. 2019. Neural cross-lingual relation extraction based on bilingual word embedding mapping. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 399–409, Hong Kong, China. Association for Computational Linguistics.

Hiroaki Ozaki, Gaku Morio, Terufumi Morishita, and Toshinori Miyoshi. 2021. Project-then-transfer: Effective two-stage cross-lingual transfer for semantic dependency parsing. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2586–2594, Online. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling Relations and Their Mentions without Labeled Text. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD '10)*.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

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 Barnabás Póczos. 2021. Re-tacred: Addressing shortcomings of the TACRED dataset. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13843–13850. AAAI Press.

Patrick Verga, David Belanger, Emma Strubell, Benjamin Roth, and Andrew McCallum. 2016. Multilingual relation extraction using compositional universal schema. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 886–896, San Diego, California. Association for Computational Linguistics.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. Technical report, Linguistic Data Consortium.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le

Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Chenhao Xie, Jiaqing Liang, Jingping Liu, Chengsong Huang, Wenhao Huang, and Yanghua Xiao. 2021. Revisiting the negative data of distantly supervised relation extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3572–3581, Online. Association for Computational Linguistics.

Omry Yadan. 2019. Hydra - a framework for elegantly configuring complex applications. Github.

Mahsa Yarmohammadi, Shijie Wu, Marc Marone, Haoran Xu, Seth Ebner, Guanghui Qin, Yunmo Chen, Jialiang Guo, Craig Harman, Kenton Murray, Aaron Steven White, Mark Dredze, and Benjamin Van Durme. 2021. Everything is all it takes: A multipronged strategy for zero-shot cross-lingual information extraction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1950–1967, Online and Punta Cana, Dominican Republic. 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.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168, Online only. Association for Computational Linguistics.

Bowei Zou, Zengzhuang Xu, Yu Hong, and Guodong Zhou. 2018. Adversarial feature adaptation for cross-lingual relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 437–448, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

A Translation Details

We use the following parameter settings for DeepL API calls: *split_sentences:1, tag_handling:xml, outline_detection:0*. For Google, we use *format_:html, model:nmt*.

Table 6 shows the number of syntactically valid and invalid translations for each language and split, as well as for the back-translation of the test split.

For tokenization, we use Spacy 3.2[14] with standard (non-neural) models for *de, es, fr, fi, ja, pl, ru, zh*, and TranKIT 1.1.0[15] for *ar, hi, hu, tr*.

The translation costs per language amount to approximately 460 Euro, for a total character count of 22.9 million characters to be translated (source sentences including entity markup tags), at a price of 20 Euro per 1 million characters at the time of writing. Compared to an estimated annotation cost of approximately 10K USD, translation costs amount to less than 5% of the cost of fully annotating a similar-sized dataset in a new language.[16]

B Human Translation Analysis

For the manual analysis of translated TACRED instances, we recruited a single native speaker for each language among the members of our lab and associated partners. Annotators were not paid for the task, but performed it as part of their work at the lab. All annotators are either Master's degree or PhD students, with a background in Linguistics, Computer Science, or a related field. The full instructions given to annotators, after a personal introduction to the task, are shown in Figure 2.

C Additional Training Details

All pre-trained models evaluated in this study are used as they are available from HuggingFace's model hub, without any modifications. Our implementation uses HF's *BertForSequenceClassification* implementation with default settings for dropout, positional embeddings, etc. Licenses for the pretrained BERT models are listed in Table 7, if specified in the repository. The Transformers library is available under the Apache 2.0 license,

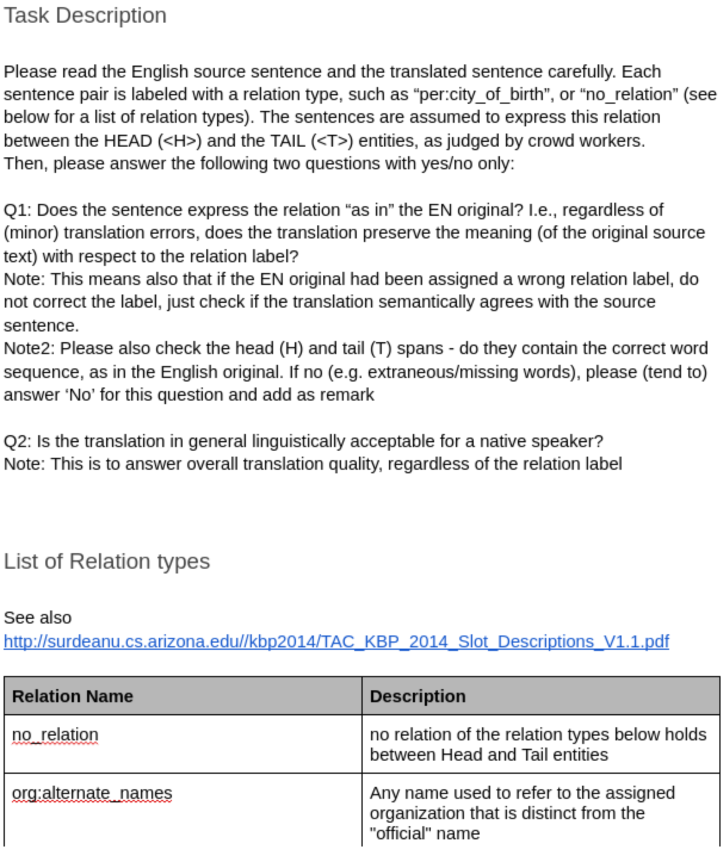
[14] https://spacy.io

[15] https://github.com/nlp-uoregon/trankit

[16] Stoica et al. (2021) pay 0.15 USD per HIT of 5 sentences in *TACRED*. With an average of 3 crowd workers per HIT and a total of 106,264 examples in *TACRED*, this amounts to approximately 9,564 USD. Angeli et al. (2014) report a cost of 3,156 USD for annotating 23,725 examples, which would correspond to a cost of 14,135 USD for the whole *TACRED* dataset.

Language (Translation Engine)	Train	Train Err	Dev	Dev Err	Test	Test Err	BT Test	BT Test Err
en (-)	68,124	-	22,631	-	15,509	-	-	-
ar (G)	67,736	388	22,502	129	15,425	84	15,425	0
de (D)	67,253	871	22,343	288	15,282	227	15,079	203
es (D)	65,247	2,877	21,697	934	14,908	601	14,688	220
fi (D)	66,751	1,373	22,268	363	15,083	426	14,462	621
fr (D)	66,856	1,268	22,298	333	15,237	272	15,088	149
hi (G)	67,751	373	22,511	120	15,440	69	15,440	0
hu (G)	67,766	358	22,519	112	15,436	73	15,436	0
ja (D)	61,571	6,553	20,290	2,341	13,701	1,808	12,913	805
pl (G)	68,124	0	22,631	0	15,509	0	15,509	0
ru (D)	66,413	1,711	21,998	633	14,995	514	14,703	292
tr (G)	67,749	375	22,510	121	15,429	80	15,429	0
zh (D)	65,260	2,864	21,538	1,093	14,694	815	14,021	681
$\cap_{all}$	54,251	-	17,809	-	11,874	-	9,944	-

Table 6: *MultiTACRED* instances per language and split, and for the back-translation (BT) of the test split. The *'en'* row shows the statistics of the original *TACRED.* (G) and (D) refer to Google and DeepL, respectively. The error columns list the number of instances discarded after translation due to missing / erroneous entity tag markup. On average, 2.3% of the instances were discarded due to invalid entity markup after translation. The last row shows the intersection of valid instances available in all languages.



Task Description

Please read the English source sentence and the translated sentence carefully. Each sentence pair is labeled with a relation type, such as “per:city_of_birth”, or “no_relation” (see below for a list of relation types). The sentences are assumed to express this relation between the HEAD (<H>) and the TAIL (<T>) entities, as judged by crowd workers.
Then, please answer the following two questions with yes/no only:

Q1: Does the sentence express the relation “as in” the EN original? I.e., regardless of (minor) translation errors, does the translation preserve the meaning (of the original source text) with respect to the relation label?
Note: This means also that if the EN original had been assigned a wrong relation label, do not correct the label, just check if the translation semantically agrees with the source sentence.
Note2: Please also check the head (H) and tail (T) spans - do they contain the correct word sequence, as in the English original. If no (e.g. extraneous/missing words), please (tend to) answer ‘No’ for this question and add as remark

Q2: Is the translation in general linguistically acceptable for a native speaker?
Note: This is to answer overall translation quality, regardless of the relation label

List of Relation types

See also
http://surdeanu.cs.arizona.edu//kbp2014/TAC_KBP_2014_Slot_Descriptions_V1.1.pdf

Relation Name	Description
no_relation	no relation of the relation types below holds between Head and Tail entities
org:alternate_names	Any name used to refer to the assigned organization that is distinct from the "official" name

Figure 2: Task description given to human judges for translation quality analysis.

Hydra under the MIT license, and PyTorch uses a modified BSD license.

For Hungarian, we use *bert-base-multilingual-cased*, since there is no pretrained Hungarian BERT model available on the hub. For Hindi, we tried several models by l3cube-pune, neuralspace-reverie, google and ai4bharat, but all of these produced far worse results than the ones reported here for *l3cube-pune/hindi-bert-scratch*. Interestingly, using *bert-base-multilingual-cased* instead of *l3cube-pune/hindi-bert-scratch* as the base PLM produced far better results for Hindi in the monolingual setting, at 71.1 micro-F1.

We experimented with learning rates in $[3e-6, 7e-6, 1e-5, 3-e5, 5e-5]$. We used micro-F1 on the *dev* set as the criterion for hyperparameter selection. Table 7 lists the best learning rates per language and scenario. We use a fixed set of random seeds {1337, 2674, 4011, 5348, 6685} for training across the 5 runs.

D Translation Error Examples

Table 8 lists common error types we identified in the translations of *TACRED* instances.

Language/Scenario	HuggingFace Model name	LR	License
ar	aubmindlab/bert-base-arabertv02	1e-5	N/A
de	bert-base-german-cased	3e-5	MIT
en	bert-base-uncased	3e-5	Apache 2.0
es	dccuchile/bert-base-spanish-wwm-cased	1e-5	(CC BY 4.0)
fi	TurkuNLP/bert-base-finnish-cased-v1	7e-6	N/A
fr	flaubert/flaubert_base_cased	1e-5	MIT
hi	l3cube-pune/hindi-bert-scratch	7e-6	CC BY 4.0
hu	bert-base-multilingual-cased	1e-5	Apache 2.0
ja	cl-tohoku/bert-base-japanese-whole-word-masking	3e-5	CC BY 4.0
pl	dkleczek/bert-base-polish-cased-v1	7e-6	N/A
ru	sberbank-ai/ruBert-base	3e-5	Apache 2.0
tr	dbmdz/bert-base-turkish-cased	1e-5	MIT
zh	bert-base-chinese	1e-5	N/A
Cross-lingual mBERT	bert-base-multilingual-cased	1e-5	Apache 2.0
Multilingual mBERT	bert-base-multilingual-cased	1e-5	Apache 2.0

Table 7: Best learning rate and model identifiers per language for the monolingual settings, and for the cross- and multilingual scenarios. The table also lists the model license, if it was available.

Error Type	**Source**	**Lang.**	**Translation**	**Comment**
Alignment - Missing	<H>He</H> also presided over the country 's <T>Constitutional Council</T> [. . .]	es	También presidió el <T>Consejo Constitucional</T> del país [. . .]	Head not marked due to dropped pronoun
Alignment - Definite Article	<T>JetBlue Airways Corp</T> spokesman <H>Bryan Baldwin</H> said [. . .]	es	<H>El</H> portavoz de<T>JetBlue Airways Corp</T> <H>, Bryan Baldwin</H>, dijo [. . .]	'El' is marked as additional head span
Alignment - Split span	New <T>York-based Human Rights Watch</T> (HRW) , [. . .] snubbed an invitation to testify [. . .]	es	<T>Human Rights Watch</T> (HRW), con sede en Nueva <T>York</T>, [. . .] rechazaron una invitación para testificar [. . .]	Translation of 'York-based' syntactically different, leading to split span
Alignment - Split Compound	[. . .] Russian <T>Foreign Ministry</T> spokesman Andrei Nesterenko said on Thursday , <H>RIA Novosti</H> reported.	fr	[. . .] a déclaré jeudi le porte-parole du <T>ministère</T> russe <T>des affaires étrangères</T>, Andrei Nesterenko, selon <H>RIA Novosti</H>.	French word order for adjectives leads to split span of compound 'Foreign Ministry'
Alignment - Compound	[. . .] Seethapathy Chander , Deputy Director General with <T>ADB</T> 's <H>Private Sector Department</H>.	de	[. . .] Seethapathy Chander, stellvertretender Generaldirektor der <H>ADB-Abteilung für den Privatsektor</H>.	German translation uses a compound noun combining head and 'department'
Alignment - Missing	<H>She</H> was vibrant , she loved life and <T>she</T> always had a kind word for everyone.	de	<H>Sie</H> war lebhaft, sie liebte das Leben und hatte immer ein freundliches Wort für jeden.	Multiple occurrences of same pronoun seem to confuse aligner
Alignment - Coordination	<H>Christopher Bentley</H> , a spokesman for Citizenship and <T>Immigration Services</T> [. . .]	es	<H>Christopher Bentley</H>, un portavoz de <T>los Servicios de</T> Ciudadanía e <T>Inmigración</T> [. . .]	Coordinated conjuction in Spanish leads to split span
Alignment - Wrong	She said when <H>she</H> got pregnant in <T>2008</T> [. . .]	pl	Powiedziała, że kiedy w <T>2008</T> r. <H>zaszła</H> w ciążę [. . .]	'got' marked instead of dropped pronoun 'she'
Alignment - Extended	<T>Alaskans</T> last chose a Democrat for the presidency in 1964 , when they backed Lyndon B. Johnson by a 2-1 margin over <H>Barry Goldwater</H> .	zh	<T>阿拉斯加人上</T>一次选择民主党人担任总统是在1964年，当时他们以2比1的优势支持林登-B-约翰逊，而不是<H>巴里-戈德华特</H>。	'last' is included in tail span
Alignment - Partial	In August , <H>Baldino</H> [. . .] had taken a leave of absence from his posts as Cephalon 's chairman and <T>chief executive</T> .	pl	W sierpniu <H>Baldino</H> [. . .] wziął urlop od pełnienia funkcji prezesa i <T>dyrektora general</T> nego firmy Cephalon.	'nego' should be part of the tail span and not be split off of the word 'generalnego'
Alignment - Inflection	Some of the people profiled are <T>ABC</T> president <H>Steve McPherson</H> , [. . .]	fi	Mukana ovat muun muassa <T>ABC:n</T> pääjohtaja <H>Steve McPherson</H>, [. . .]	Tail 'ABC:n' includes genitive case marker in Finnish
Non-English Source	Dari arah Jakarta/Indramayu , <T>sekitar</T> 2 km sebelum Pasar Celancang , tepatnya di sebelah Kantor Kecamatan Suranenggala terdapat Tempat Pelelangan Ikan (<H>TPI</H>) .	-	-	Source language is Indonesian, not English
Sentence split	<H>Stewart</H> is not saying that a 1987-style stock market crash is on the immediate horizon , and <T>he</T> concedes that " by many measures , stocks are n't overpriced , even at recent highs . "	tr	<H>Stewart</H>, 1987 tarzı bir borsa çöküşünün hemen ufukta olduğunu söylemiyor ve <T>o</T> " birçok önlemle , hisse senetlerinin aşırı fiyatlandırılmadığını bile kabul ediyor . son zirvelerde. "	'son zirvelerde' erroneously separated by end-of-sentence period
Translation incomplete	Outlined in a filing with the <H>Federal Election Commission</H> , <T>Obama</T> 's suggestion is notable because . . .	de	Der Vorschlag <T>Obamas</T> ist bemerkenswert, weil . . .	Translation is missing first part and head span
Atypical input	Browns 5-10 [. . .] <T>Cowboys</T> 5-10 [. . .] <H>Jaguars</H> 8-7 [. . .] Total : 42-93 (.311) Total : 58-74 (.439) Total : 53-81 (.396)	zh	Browns 5-10 [. . .] <T>Cowboys</T>5-10 [. . .] <H>Jaguars</H>8-7 [. . .] Total : 42-93 (.311) 总数: 58-74 (.439) 总数: 53-81 (.396)	Almost no translation due to atypical input

Table 8: Common error types of translated *TACRED* examples. The first half of the table shows alignment errors that can be automatically detected, such as missing or additional aligned spans in the translation. The second half shows error types identified by human judges.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations Section

☑ A2. Did you discuss any potential risks of your work?
Section Limitations & Section Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Sections 2.1 TACRED, 2.2 Translation Systems, 3.2 Models/Libraries, Appendix A Preprocessing

☑ B1. Did you cite the creators of artifacts you used?
2.1 TACRED, 2.2 Translation Systems, 3.2 Models/Libraries, Appendix A Preprocessing

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
2.1 TACRED; 2.2 Translation Systems, Appendix C Models/Libraries

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
TACRED - Section 2, MultiTacred - Section Conclusion

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
TACRED - Sec Ethics Statement

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2.1 & 2.2

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
TACRED/MultiTACRED - Table 6

C ☑ Did you run computational experiments?

Section 3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Section 3, Ethics Statement & Appendix C

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 3, Appendix C

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3.2

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix A Preprocessing

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Section 2.3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix B

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix B

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Explained informally during introduction to the task

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
not applicable

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Appendix B