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; Khalidi 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.\footnote{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).} We select typologically diverse languages which span a large set of linguistic phenomena such as compounding, inflection and pronoun-drop,
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 monolingual 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

---

2https://catalog.ldc.upenn.edu/LDC2018T24, under a LDC license
4The first row of Table 6 in Appendix A summarizes key statistics of the dataset.
5https://api.deepl.com/v2/translate
6https://translation.googleapis.com/language/translate/v3
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. 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 Multi-TACRED 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? (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

---

7See Appendix A for details.

---
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 \( R \) and a text \( x = [x_1, x_2, \ldots, x_n] \) (where \( x_1, \ldots, x_n \) are tokens) with two disjoint spans \( e_h = [x_i, \ldots, x_j] \) and \( e_t = [x_k, \ldots, x_l] \) denoting the head and tail entity mentions, RE aims to predict the relation \( r \in R \) between \( e_h \) and \( e_t \), or assign the no_relation class if no relation in \( R \) 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). 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]”. 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, \( \beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e-8 \), 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.\(^9\) 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\(_I\)). 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). 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.

### Table 1: Translation quality, as judged by native speakers.

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

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. 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. Their morpheme level, but this in turn may raise issues with e.g. PLM tokenization and entity masking.

---

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...
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 \((test_\cap)\).

<table>
<thead>
<tr>
<th>Test set</th>
<th>en</th>
<th>ar</th>
<th>de</th>
<th>es</th>
<th>fi</th>
<th>fr</th>
<th>hi</th>
<th>hu</th>
<th>ja</th>
<th>pl</th>
<th>ru</th>
<th>tr</th>
<th>zh</th>
</tr>
</thead>
<tbody>
<tr>
<td>(test_L)</td>
<td>77.1</td>
<td>74.2</td>
<td>74.1</td>
<td>75.7</td>
<td>76.4</td>
<td>75.0</td>
<td>65.1</td>
<td>71.8</td>
<td>71.8</td>
<td>73.7</td>
<td>73.7</td>
<td>74.2</td>
<td>75.4</td>
</tr>
<tr>
<td>(test_\cap)</td>
<td>77.5</td>
<td>74.5</td>
<td>74.6</td>
<td>76.1</td>
<td>76.6</td>
<td>75.4</td>
<td>65.9</td>
<td>72.4</td>
<td>72.5</td>
<td>74.3</td>
<td>74.8</td>
<td>74.5</td>
<td>75.3</td>
</tr>
</tbody>
</table>

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.
<table>
<thead>
<tr>
<th>Test set / WikiSize</th>
<th>Metric</th>
<th>en</th>
<th>ar</th>
<th>de</th>
<th>es</th>
<th>fi</th>
<th>fr</th>
<th>hi</th>
<th>hu</th>
<th>ja</th>
<th>pl</th>
<th>ru</th>
<th>tr</th>
<th>zh</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>76.7</td>
<td>72.1</td>
<td>75.2</td>
<td>74.0</td>
<td>76.7</td>
<td>74.3</td>
<td>76.1</td>
<td>76.5</td>
<td>78.6</td>
<td>76.9</td>
<td>70.6</td>
<td>73.6</td>
<td>73.2</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>77.5</td>
<td>60.3</td>
<td>74.0</td>
<td>73.9</td>
<td>64.9</td>
<td>73.9</td>
<td>53.0</td>
<td>59.7</td>
<td>51.3</td>
<td>70.0</td>
<td>74.6</td>
<td>57.4</td>
<td>70.0</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>77.1</td>
<td>65.7</td>
<td>74.6</td>
<td>73.9</td>
<td>70.3</td>
<td>74.1</td>
<td>62.5</td>
<td>67.1</td>
<td>62.1</td>
<td>73.3</td>
<td>72.6</td>
<td>64.5</td>
<td>71.6</td>
</tr>
<tr>
<td></td>
<td>P</td>
<td>76.5</td>
<td>73.2</td>
<td>75.5</td>
<td>74.8</td>
<td>78.3</td>
<td>75.0</td>
<td>76.5</td>
<td>76.6</td>
<td>79.2</td>
<td>77.1</td>
<td>70.6</td>
<td>73.4</td>
<td>73.8</td>
</tr>
<tr>
<td></td>
<td>R</td>
<td>78.3</td>
<td>61.6</td>
<td>74.3</td>
<td>75.3</td>
<td>65.1</td>
<td>74.3</td>
<td>54.3</td>
<td>60.7</td>
<td>50.8</td>
<td>71.1</td>
<td>75.3</td>
<td>58.1</td>
<td>69.9</td>
</tr>
<tr>
<td></td>
<td>F1</td>
<td>77.4</td>
<td>66.9</td>
<td>74.9</td>
<td>75.0</td>
<td>71.1</td>
<td>74.6</td>
<td>63.5</td>
<td>67.7</td>
<td>61.9</td>
<td>74.0</td>
<td>72.9</td>
<td>64.9</td>
<td>71.8</td>
</tr>
</tbody>
</table>

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**<sub>L</sub>) and on the intersection of test instances available in all languages (**test**<sub>∩</sub>), 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<sub>2</sub>(MegaBytes), taken from Wu and Dredze (2020)). Languages with less pretraining data in mBERT suffer a larger performance loss.

<table>
<thead>
<tr>
<th>WikiSize log&lt;sub&gt;2&lt;/sub&gt;(MB)</th>
<th>14</th>
<th>10</th>
<th>12</th>
<th>12</th>
<th>9</th>
<th>12</th>
<th>7</th>
<th>10</th>
<th>11</th>
<th>11</th>
<th>12</th>
<th>9</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-lang data (%)</td>
<td>ar</td>
<td>de</td>
<td>es</td>
<td>fi</td>
<td>fr</td>
<td>hi</td>
<td>hu</td>
<td>ja</td>
<td>pl</td>
<td>ru</td>
<td>tr</td>
<td>zh</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
</tr>
<tr>
<td>5</td>
<td>65.7</td>
<td>74.6</td>
<td>73.9</td>
<td>70.3</td>
<td>74.1</td>
<td>62.5</td>
<td>67.1</td>
<td>62.1</td>
<td>73.3</td>
<td>72.6</td>
<td>64.5</td>
<td>71.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>68.7</td>
<td>74.9</td>
<td>73.9</td>
<td>70.5</td>
<td>74.3</td>
<td>67.0</td>
<td>68.8</td>
<td>69.2</td>
<td>72.2</td>
<td>73.0</td>
<td>67.8</td>
<td>72.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>69.0</td>
<td>74.5</td>
<td>73.7</td>
<td>70.4</td>
<td>73.6</td>
<td>68.0</td>
<td>68.9</td>
<td>70.6</td>
<td>72.0</td>
<td>72.7</td>
<td>68.9</td>
<td>73.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>71.0</td>
<td>74.4</td>
<td>74.5</td>
<td>72.2</td>
<td>73.9</td>
<td>69.9</td>
<td>70.2</td>
<td>71.7</td>
<td>73.3</td>
<td>73.3</td>
<td>69.2</td>
<td>73.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>71.4</td>
<td>74.8</td>
<td>74.8</td>
<td>72.3</td>
<td>74.2</td>
<td>70.1</td>
<td>71.0</td>
<td>72.3</td>
<td>72.9</td>
<td>73.2</td>
<td>70.1</td>
<td>73.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>71.2</td>
<td>74.3</td>
<td>74.5</td>
<td>72.1</td>
<td>73.9</td>
<td>70.4</td>
<td>70.8</td>
<td>71.6</td>
<td>73.0</td>
<td>73.1</td>
<td>70.3</td>
<td>74.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>100</td>
<td>73.5</td>
<td>75.8</td>
<td>75.9</td>
<td>73.5</td>
<td>75.6</td>
<td>72.4</td>
<td>72.4</td>
<td>73.5</td>
<td>74.3</td>
<td>75.6</td>
<td>71.6</td>
<td>75.4</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 Özugr, 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 (Khalidi 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,
### 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

<table>
<thead>
<tr>
<th>Language</th>
<th>ar</th>
<th>de</th>
<th>es</th>
<th>fi</th>
<th>fr</th>
<th>hi</th>
<th>hu</th>
<th>ja</th>
<th>pl</th>
<th>ru</th>
<th>tr</th>
<th>zh</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>69.6</td>
<td>76.1</td>
<td>75.8</td>
<td>73.6</td>
<td>75.9</td>
<td>73.3</td>
<td>70.0</td>
<td>72.2</td>
<td>74.7</td>
<td>74.0</td>
<td>72.1</td>
<td>74.8</td>
</tr>
</tbody>
</table>

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.
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


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


Talaat Khalil, Kornel Kielczewski, Georgios Christos Choularias, Amina Keldibek, and Maarten Versteegh.


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


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 3.214 with standard (non-neural) models for de, es, fr, fi, ja, pl, ru, zh, and TranKIT 1.1.015 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.

14https://spacy.io
15https://github.com/nlp-uoregon/trankit
16Stoica 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.
<table>
<thead>
<tr>
<th>Language (Translation Engine)</th>
<th>Train</th>
<th>Train Err</th>
<th>Dev</th>
<th>Dev Err</th>
<th>Test</th>
<th>Test Err</th>
<th>BT Test</th>
<th>Test Err</th>
<th>BT Test</th>
<th>BT Test Err</th>
</tr>
</thead>
<tbody>
<tr>
<td>en (-)</td>
<td>68,124</td>
<td>-</td>
<td>22,631</td>
<td>15,509</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>ar (G)</td>
<td>67,736</td>
<td>388</td>
<td>22,502</td>
<td>15,425</td>
<td>84</td>
<td>15,425</td>
<td>0</td>
<td>203</td>
<td></td>
<td></td>
</tr>
<tr>
<td>de (D)</td>
<td>67,253</td>
<td>871</td>
<td>22,343</td>
<td>15,282</td>
<td>227</td>
<td>15,079</td>
<td>230</td>
<td>149</td>
<td></td>
<td></td>
</tr>
<tr>
<td>es (D)</td>
<td>65,247</td>
<td>2,877</td>
<td>21,697</td>
<td>14,908</td>
<td>601</td>
<td>14,688</td>
<td>220</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>fa (D)</td>
<td>66,751</td>
<td>1,373</td>
<td>22,268</td>
<td>15,425</td>
<td>84</td>
<td>15,425</td>
<td>0</td>
<td>220</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>fr (D)</td>
<td>66,856</td>
<td>2,877</td>
<td>22,298</td>
<td>15,237</td>
<td>272</td>
<td>15,088</td>
<td>149</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>hi (G)</td>
<td>67,751</td>
<td>373</td>
<td>22,511</td>
<td>15,440</td>
<td>69</td>
<td>15,440</td>
<td>0</td>
<td>203</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hu (G)</td>
<td>67,766</td>
<td>358</td>
<td>22,519</td>
<td>15,436</td>
<td>73</td>
<td>15,436</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ja (D)</td>
<td>61,571</td>
<td>6,553</td>
<td>20,290</td>
<td>13,730</td>
<td>1,808</td>
<td>12,913</td>
<td>805</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pl (G)</td>
<td>68,124</td>
<td>0</td>
<td>22,631</td>
<td>15,509</td>
<td>0</td>
<td>15,509</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ru (D)</td>
<td>66,413</td>
<td>1,711</td>
<td>21,998</td>
<td>14,995</td>
<td>514</td>
<td>14,703</td>
<td>292</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tr (G)</td>
<td>67,749</td>
<td>375</td>
<td>22,510</td>
<td>15,429</td>
<td>80</td>
<td>15,429</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>zh (D)</td>
<td>65,260</td>
<td>2,864</td>
<td>21,538</td>
<td>14,694</td>
<td>815</td>
<td>14,021</td>
<td>681</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

∩_{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 “pair 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 (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.

Note 2: 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


<table>
<thead>
<tr>
<th>Relation Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>no_relation</td>
<td>no relation of the relation types below holds between Head and Tail entities</td>
</tr>
<tr>
<td>org_alternate_names</td>
<td>Any name used to refer to the assigned organization that is distinct from the &quot;official&quot; name</td>
</tr>
</tbody>
</table>

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
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 $[3e^{-6}, 7e^{-6}, 1e^{-5}, 3e^{-5}, 5e^{-5}]$. 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.
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.

<table>
<thead>
<tr>
<th>Language/Scenario</th>
<th>HuggingFace Model name</th>
<th>LR</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td>ar</td>
<td>aubmindlab/bert-base-arabertv02</td>
<td>1e-5</td>
<td>N/A</td>
</tr>
<tr>
<td>de</td>
<td>bert-base-german-cased</td>
<td>3e-5</td>
<td>MIT</td>
</tr>
<tr>
<td>en</td>
<td>bert-base-uncased</td>
<td>3e-5</td>
<td>Apache 2.0</td>
</tr>
<tr>
<td>es</td>
<td>dccuchile/bert-base-spanish-wwn-cased</td>
<td>1e-5</td>
<td>(CC BY 4.0)</td>
</tr>
<tr>
<td>fi</td>
<td>TurkulNLP/bert-base-finnish-cased-v1</td>
<td>7e-6</td>
<td>N/A</td>
</tr>
<tr>
<td>fr</td>
<td>flaubert/flaubert_base_cased</td>
<td>1e-5</td>
<td>MIT</td>
</tr>
<tr>
<td>hi</td>
<td>i3cube-pune/hindi-bert-scratch</td>
<td>7e-6</td>
<td>CC BY 4.0</td>
</tr>
<tr>
<td>hu</td>
<td>bert-base-multilingual-cased</td>
<td>1e-5</td>
<td>Apache 2.0</td>
</tr>
<tr>
<td>ja</td>
<td>cl-tohoku/bert-base-japanese-whole-word-masking</td>
<td>3e-5</td>
<td>CC BY 4.0</td>
</tr>
<tr>
<td>pl</td>
<td>dklczech/bert-base-polish-cased-v1</td>
<td>7e-6</td>
<td>N/A</td>
</tr>
<tr>
<td>ru</td>
<td>sberbank-aiuruBert-base</td>
<td>3e-5</td>
<td>Apache 2.0</td>
</tr>
<tr>
<td>tr</td>
<td>dbmdz/bert-base-turkish-cased</td>
<td>1e-5</td>
<td>MIT</td>
</tr>
<tr>
<td>zh</td>
<td>bert-base-chinese</td>
<td>1e-5</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Cross-lingual mBERT bert-base-multilingual-cased 1e-5 Apache 2.0

Multilingual mBERT bert-base-multilingual-cased 1e-5 Apache 2.0

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.