Multilingual Event Extraction from Historical Newspaper Adverts

WARNING: This paper shows dataset samples which are racist in nature

Nadav Borenstein
University of Copenhagen
nadav.borenstein@di.ku.dk

Natália da Silva Perez
Erasmus University Rotterdam
dasilvaperez@eshcc.eur.nl

Isabelle Augenstein
University of Copenhagen
augenstein@di.ku.dk

Abstract

NLP methods can aid historians in analyzing textual materials in greater volumes than manually feasible. Developing such methods poses substantial challenges though. First, acquiring large, annotated historical datasets is difficult, as only domain experts can reliably label them. Second, most available off-the-shelf NLP models are trained on modern language texts, rendering them significantly less effective when applied to historical corpora. This is particularly problematic for less well studied tasks, and for languages other than English. This paper addresses these challenges while focusing on the under-explored task of event extraction from a novel domain of historical texts. We introduce a new multilingual dataset in English, French, and Dutch composed of newspaper ads from the early modern colonial period reporting on enslaved people who liberated themselves from enslavement. We find that: 1) even with scarce annotated data, it is possible to achieve surprisingly good results by formulating the problem as an extractive QA task and leveraging existing datasets and models for modern languages; and 2) cross-lingual low-resource learning for historical languages is highly challenging, and machine translation of the historical datasets to the considered target languages is, in practice, often the best-performing solution.

1 Introduction

Analyzing large corpora of historical documents can provide invaluable insights on past events in multiple resolutions, from the life of an individual to processes on a global scale (Borenstein et al., 2023; Laite, 2020; Gerritsen, 2012). While historians traditionally work closely with the texts they study, automating parts of the analysis using NLP tools can help speed up the research process and facilitate the extraction of historical evidence from large corpora, allowing historians to focus on interpretation.

However, building NLP models for historical texts poses a substantial challenge. First, acquiring large, annotated historical datasets is difficult (Hämäläinen et al., 2021; Bollmann and Søgaard, 2016), as only domain experts can reliably label them. This renders the default fully-supervised learning setting less feasible for historical corpora. Compounding this, most off-the-shelf NLP models were trained on modern language texts and display significantly weaker performance for historical documents (Manjavacas and Fonteyn, 2022; Baptiste et al., 2021; Hardmeier, 2016), which usually suffer from a high rate of OCR errors and are written in a substantially different language. This is particularly challenging for less well-studied tasks or for non-English languages.

One of these under-explored tasks is event extraction from historical texts (Sprugnoli and Tonelli, 2019; Lai et al., 2021), which can aid in retrieving information about complex events from vast amounts of texts. Here, we research extraction of events from adverts in colonial newspapers reporting on enslaved people who escaped their enslavers. Studying these ads can shed light on the linguistic processes of racialization during the early modern colonial period (c. 1450 to 1850), the era of the transatlantic slave trade, which coincided with the early era of mass print media.

Methodologically, we research low-resource learning methods for event extraction, for which only a handful of prior papers exist (Lai et al., 2021; Sprugnoli and Tonelli, 2019). To the best of our knowledge, this is the first paper to study historical event extraction in a multilingual setting.

Specifically, our contributions are as follows:

- We construct a new multilingual dataset in English, French, and Dutch of *"freedom-seeking events"*, composed of ads placed by enslavers reporting on enslaved people who sought freedom by escaping them, building on an existing annotated English language dataset of "run-

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 10304–10325

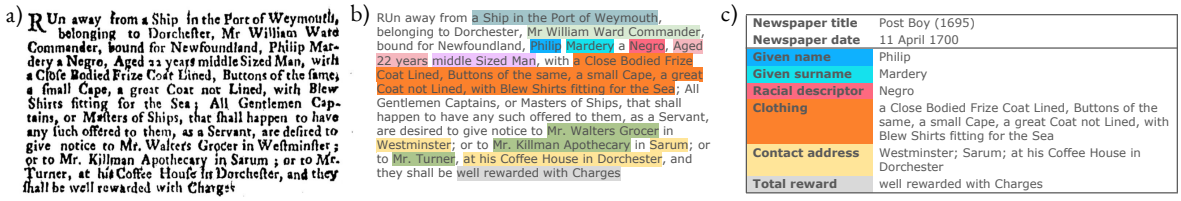


Figure 1: An example from the annotated *Runaway Slaves in Britain* dataset. Each data point includes a scan of the ad (a), the extracted text (b), and a list of attributes that appear in the ad as well as relevant metadata (c).

away slave adverts" (Newman et al., 2019).[1] Fig. 1a contains an example ad.

- We propose to frame event extraction from historical texts as extractive question answering. We show that even with scarce annotated data, this formulation can achieve surprisingly good results by leveraging existing resources for modern languages.
- We show that cross-lingual low-resource learning for historical languages is highly challenging, and machine translation of the historical datasets to the target languages is often the best-performing solution in practice.

2 Related Work

2.1 NLP for Historical Texts

Prior work on NLP for historical texts has mainly focused on OCR and text normalization (Drobac et al., 2017; Robertson and Goldwater, 2018; Bollmann et al., 2018; Bollmann, 2019; Lyu et al., 2021). However, NLP has also been used to assist historians in analyzing large amounts of textual material in more complex ways. Recent work has researched tasks such as PoS tagging (Yang and Eisenstein, 2016), Named Entity Recognition (Ehrmann et al., 2021; De Toni et al., 2022) and co-reference resolution (Darling et al., 2022; Krug et al., 2015), and bias analysis (Borenstein et al., 2023). Many of these studies report the difficulties of acquiring large annotated historical datasets (Hämäläinen et al., 2021; Bollmann and Søgaard, 2016) and replicating the impressive results of large pre-trained language models on modern texts (Lai et al., 2021; De Toni et al., 2022). This also led prior work to focus on monolingual texts, particularly in English, while neglecting low-resource languages. In this paper, we attempt to alleviate these challenges while investigating a task that is underexplored from the perspective of historical NLP – multilingual event extraction.

2.2 Event Extraction

Event extraction (Hogenboom et al., 2011; Xiang and Wang, 2019) is the task of organising natural text into structured events – specific occurrences of something that happens at a particular time and place involving one or more participants, each associated with a set of attributes.

Traditionally, event extraction is decomposed into smaller, less complex subtasks (Lin et al., 2020; Li et al., 2020), such as detecting the existence of an event (Weng and Lee, 2011; Nguyen and Grishman, 2018; Sims et al., 2019), identifying its participants (Du et al., 2021; Li et al., 2020), and extracting the attributes associated with the event (Li et al., 2020; Zhang et al., 2020; Du and Cardie, 2020). Recent work (Liu et al., 2020; Du and Cardie, 2020) has shown the benefit of framing event extraction as a QA task, especially for the sub-task of attribute extraction, which is the focus of this work. We build on the latter finding, by framing the identification of attributes associated with historical events as an extractive QA task.

Event extraction from historical texts is much less well studied than extraction from modern language texts, with only a handful of works targeting this task. Cybulska and Vossen (2011); Segers et al. (2011) develop simple pipelines for extracting knowledge about historical events from modern Dutch texts. Sprugnoli and Tonelli (2019) define annotation guidelines for detecting and classifying events mentioned in historical texts and compare two models on a new corpus of historical documents. Boros et al. (2022) study the robustness of two event detection models to OCR noise by automatically degrading modern event extraction datasets in several languages. Finally, and closer to this work, Lai et al. (2021) present BRAD, a dataset for event extraction from English historical texts about Black rebellions, which is not yet

[1] We make our dataset and code publicly available at https://github.com/nadavborenstein/EE-from-historical-ads

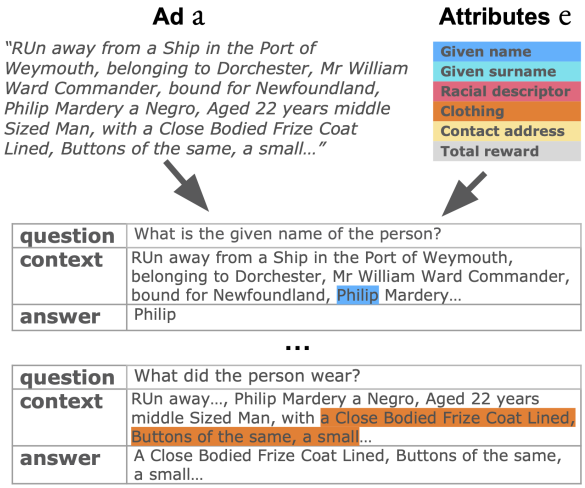


Figure 2: Our data processing pipeline: each ad is converted to a collection of extractive QA examples, where each attribute is mapped to a natural language question.

publicly available. They find that there is a significant gap in the performance of current models on BRAD compared to modern datasets. Conversely, we explore event extraction in a multilingual setting while performing a more exhaustive evaluation of various models and pipelines.

3 Methods

We now describe the methodology of the paper, including problem formulation (§3.1), datasets (§3.2), models (§3.3), and the experiments setup (§3.4).

3.1 Problem Formulation

Our starting point is a dataset where each sample is an ad corresponding to a single event. Therefore, we do not have to use event triggers – we already know what event appeared in each sample (a freedom-seeking event). We focus instead on the sub-task of attribute extraction. Following prior work (Liu et al., 2020), we formulate the problem as an extractive QA task (see Fig. 2). Specifically, given an advert a and an event attribute e, we convert e into a natural question q and search for a text span in a that answers q. We convert the attributes to questions manually;[2] see §3.2 for details. For example, if a is the attribute "total reward", we look for a text span in a that answers the question "How much reward is offered?".

We opt for this formulation for several reasons. First, extractive QA has the advantage of retrieving event attributes in the form of a span that appears verbatim in the historical document. This feature is crucial for historians, who might not trust other types of output (an abstractive QA model might generate paraphrases of the attribute or even hallucinate nonexistent facts (Zhou et al., 2021)).

Second, this formulation is especially useful in low resource settings. As annotating historical corpora is expensive and labour-intensive, these settings are prevalent in historical domains. Extractive QA is a well-researched task, with many existing datasets (Rajpurkar et al., 2016; Artetxe et al., 2019; Bartolo et al., 2020) and model checkpoints (Deepset, 2022b,a) targeting this problem. While based on modern text, the checkpoints could still be used for transfer learning (§3.3 lists the models we use for transfer learning).

Finally, an extractive QA formulation is efficient – as each event is composed of different attributes, each of which becomes a single training instance, one annotated historical ad corresponds to multiple training examples. In addition, a single model can be applied to all attribute types. This allows for a simpler and cheaper deployment, as well as a model that can benefit from multitask training and can more easily generalize to unseen attributes (§4.5).

Note that here we assume a dataset where each sample is an ad corresponding to a single self-liberation event. This setting differs from works focusing on the sub-task of event detection, e.g. using event triggers (Sims et al., 2019).

3.2 Datasets

We use a combination of annotated and unannotated datasets in three languages from different sources. See Tab. 1 for a summary of the datasets and their respective sizes.

Annotated Dataset The primary resource we use in our evaluation is an annotated English dataset scraped from the website of the *Runaways Slaves in Britain* project (Newman et al., 2019), a searchable database of over 800 newspaper adverts printed between 1700 and 1780 placed by enslavers who wanted to capture enslaved people who had self-liberated. Each ad was manually transcribed and annotated with more than 50 different attributes, such as the described gender and age, what clothes the enslaved person wore, and their physical description. See Fig. 1 for an example instance.

We clean and split the dataset into training and validation sets (70 / 30% split), and pre-process it

[2] We assume a small number of well-defined attributes of interest, as is common for historical research.

Dataset	Language	#Labeled ads	#Labeled attributes	#Unlabeled ads
Runaways Slaves in Britain	en	835	8 270	0
Runaways Slaves in Britain	fr (translated)	834	8 238	0
Runaways Slaves in Britain	nl (translated)	834	8 234	0
Marronage	en	0	0	3 026
Marronage	fr	41	313	19 066
Delpher	nl	44	272	2 742 issues

Table 1: Sizes of the different datasets.

to match the format of SQuAD-v2 (Rajpurkar et al., 2016), a large benchmark for extractive QA.[3] This involves converting each attribute into a natural language question. To find the best natural question for each attribute we first manually generate five natural questions per attribute. We then take a frozen pre-trained extractive QA model (RoBERTa-base (Liu et al., 2019) fine-tuned on SQuAD-v2) and use it to predict that attribute from the train set using each candidate question. We choose the question that results in the highest SQuAD-v2 $F1$ (Rajpurkar et al., 2018). Tab. 8 in App. D lists the resulting attributes paired with natural questions.

As no comparable datasets exist for languages other than English, we automatically translated the training split of the *Runaway Slaves in Britain* dataset into French and Dutch to support supervised training in those languages. To ensure the quality of the translation, we asked native speakers to rate 20 translations on a Likert scale of 1-5 for accuracy and fluency. Tab. 5 in App. A.2 suggests that the quality of the translations is sufficiently good. However, the translation process may have introduced a bias towards modern language, which could affect performance on these languages compared to English (§4). See App. A.2 for a description of the translation process and its evaluation.

Unannotated datasets In addition to the relatively small annotated dataset in English, we also collected an unannotated dataset of adverts in French and English scraped from *Marronage dans le monde atlantique*,[4] a platform that contains more than 20,000 manually transcribed newspaper ads about escaped enslaved people, published in French and English between the years 1765 – 1833.

For Dutch, no datasets of pre-extracted ads of such events exist yet, and we thus manually construct it. We use 2,742 full issues of the newspaper *De Curaçaosche courant*, scraped from *Delpher*,[5] a searchable API of millions of digitized OCRd texts from Dutch newspapers, books and magazines from all time periods. *De Curaçaosche courant* was chosen because almost all its issues from 1816 – 1882 are available, and it was printed mostly in Dutch (with some sections in other languages) in the Caribbean island of Curaçao, a Dutch colony during the time period we are concerned with. It is worth noting that, due to the OCR process, this dataset is noisier than the others mentioned above.

Multilingual evaluation dataset To accurately evaluate our methods on French and Dutch in addition to English, two historians of the early modern period who work with those languages manually annotated 41 and 44 adverts from the French *Marronage* and the Dutch *Delpher* corpora, respectively. As our Dutch dataset is composed of entire newspaper issues and not individual ads, the historians had first to find relevant ads before they could annotate them. The historians were guided to annotate the ads using the same attributes of the English *Runaways Slaves in Britain* dataset. See App. B for annotation guidelines.

Due to the expertise of the annotators and the annotation process being highly time-consuming, most ads were annotated by a single historian. Additionally, a random sample of 15 ads per language was annotated by a second annotator to calculate inter-annotator agreement (IAA) and assess the task's difficulty. The pairwise $F1$ agreement score (Tang et al., 2021) for each language is calculated using the 15 dual-annotated ads, yielding high $F1$ scores of 91.5, 83.2 and 80.7 for English, French and Dutch respectively. The higher agreement rate for English might be attributed to the cleaner source material in that language and possible differences in the complexity of the sources.

In summary, we now have annotated datasets in

[3] We had to discard some attributes and annotations as the annotations did not always appear verbatim in the adverts and, in some cases, could not be mapped back to the ads.

[4] www.marronnage.info/fr/index.html

[5] www.delpher.nl

three languages – the *Runaway Slaves in Britain* in English randomly divided into train and validation splits, train sets in French and Dutch generated by translating the English train set, and manually annotated validation sets in French and Dutch.

3.3 Models

Ours We experimented with several models trained with an extractive QA objective (see App. A.4 for hyper-parameters) and evaluated them using the standard SQuAD-v2 $F1$ metric. We use standard RoBERTa-based monolingual models to be evaluated in monolingual settings, as it is a well-researched model known to achieve good performance on many downstream tasks and is available in English (RoBERTa), French (CamemBERT; Martin et al., 2020) and Dutch (RobBERT; Delobelle et al., 2020). We also test variations of these models, available in English, French and Dutch, that were successively fine-tuned on large extractive QA datasets. The English models were fine-tuned on SQuAD-v2, whereas the French models were fine-tuned on a collection of three datasets – PIAF-v1.1 (Etalab, 2021), FQuAD (d'Hoffschmidt et al., 2020) and SQuAD-FR (Kabbadj, 2021). The Dutch model was fine-tuned on SQuAD-NL, a machine-translated version of SQuAD-v2.[6] In addition, we evaluate multilingual models of the XLM-RoBERTa (Conneau et al., 2019) family. We also test a variation of these models fine-tuned on SQuAD-v2. Finally, we investigate language models pre-trained on historical textual material, which are potentially better equipped to deal with historical ads. Specifically, we analyze the performance of MacBERTh (Manjavacas and Fonteyn, 2022), a BERT-based model (Devlin et al., 2019) that was pre-trained on historical textual material in English from 1450 to 1950. We also evaluate BERT models in English, French, and Dutch (Schweter, 2020, 2021a,b) that were trained specifically on historical newspapers from the 18th and the 19th centuries. Similarly, we also test variants of these models that were later fine-tuned on SQuAD.

Baselines We compare our models to two baselines suggested in prior work. De Toni et al. (2022) used a T0++ model (Sanh et al., 2021), an encoder-decoder transformer with strong zero-shot capabilities, to perform NER tagging with historical texts in several languages. We adapt this to our task by converting the evaluation examples into prompts and feeding them into T0++ (See App. A.3 for additional details). We also compare to OneIE (Lin et al., 2020), an English-only event extraction framework proposed by Lai et al. (2021).

Recall that Liu et al. (2020) also constructed event extraction as a QA task. However, their model cannot be directly compared to ours – Liu et al. supports only single sentences, while we process entire paragraphs; and adapting their model to new events which do not appear in their training dataset (as in our case) would require extensive effort, specifically for the multilingual settings. We thus leave such an investigation for future work.

3.4 Experimental Setup

The main goal of this paper is to determine the most successful approach for event extraction from historical texts with varying resources (e.g. the number of annotated examples or the existence of datasets in various languages). We therefore evaluate the models described in §3.3 with the following settings.

Zero-shot inference This simulates the prevalent case for historical NLP where no in-domain data is available for training.

Few-shot training Another frequent setup in the historical domain is where experts labeled a small number of training examples. Therefore, we train the models on our annotated monolingual datasets of various sizes (from a few examples to the entire dataset) and test their performance on evaluation sets in the same language.

Semi-supervised training Sometimes, in addition to a few labeled examples, a larger unlabeled dataset is available. We thus also evaluate our monolingual models in semi-supervised settings, where we either: 1) further pre-train the models with a masked language modeling objective (MLM) using the unannotated dataset, then fine-tune them on our annotated dataset; 2) simultaneously train the models with an MLM objective using the unannotated dataset and on the standard QA objective using the annotated dataset; or 3) use an iterative tri-training (Zhou and Li, 2005) setup to utilize the larger unannotated dataset. In tri-training, three models are trained on a labeled dataset and are used to predict the labels of unlabeled examples. All the samples for which at least two models agree on are added to the labeled set. Finally, a new model is trained on the resulting larger labeled dataset.

[6] We translated it following the procedure described in (Kabbadj, 2021).

Model	Fine-tune data	$F1$
en		
OneIE	N\A	51.90
T0++	N\A	33.69
RoBERTa-base	SQuAD-v2	54.35
RoBERTa-large	SQuAD-v2	**56.42**
XLM-RoBERTa-base	SQuAD-v2	41.84
XLM-RoBERTa-large	SQuAD-v2	55.10
fr		
T0++	N\A	32.26
CamemBERT-base	PIAF-v1.1 FQuAD-v1 SQuAD-FR	30.65
XLM-RoBERTa-base	SQuAD-v2	36.51
XLM-RoBERTa-large	SQuAD-v2	**44.52**
nl		
T0++	N\A	29.28
RobBERT-base	SQuAD-NL	37.21
XLM-RoBERTa-base	SQuAD-v2	37.56
XLM-RoBERTa-large	SQuAD-v2	**40.42**

Table 2: Zero-shot performance of different models.

Cross-lingual training Finally, we test two cross-lingual training variations. In the simple setting, we train a multilingual model on the labeled English dataset, evaluating it on French or Dutch. In the MLM settings, we also train the model with an MLM objective using the unlabeled target data.

4 Results and Analysis

4.1 Zero-Shot Inference

Tab. 2 demonstrates the benefit of framing event extraction as extractive QA. Indeed, almost all the QA models outperform the T0++ baseline by a large margin. Most English models also have significant gains over OneIE. As can also be observed from the table, the overall performance is much better for English compared to Dutch and French. This performance gap can likely be attributed to differences in the sources from which the datasets were curated. The higher IAA for the English dataset (§3.2) further supports this hypothesis. In addition, since English is the most high-resource language (Wu and Dredze, 2020), models trained on it are expected to perform best. This difference in availability of resources might also explain why the multilingual models perform better than the monolingual models on French and Dutch, while the monolingual models outperform the multilingual ones for English (Rust et al., 2021). Unsurprisingly, it can also be seen that the larger LMs achieve significantly higher $F1$ scores compared to the smaller models.

4.2 Few-Shot Training

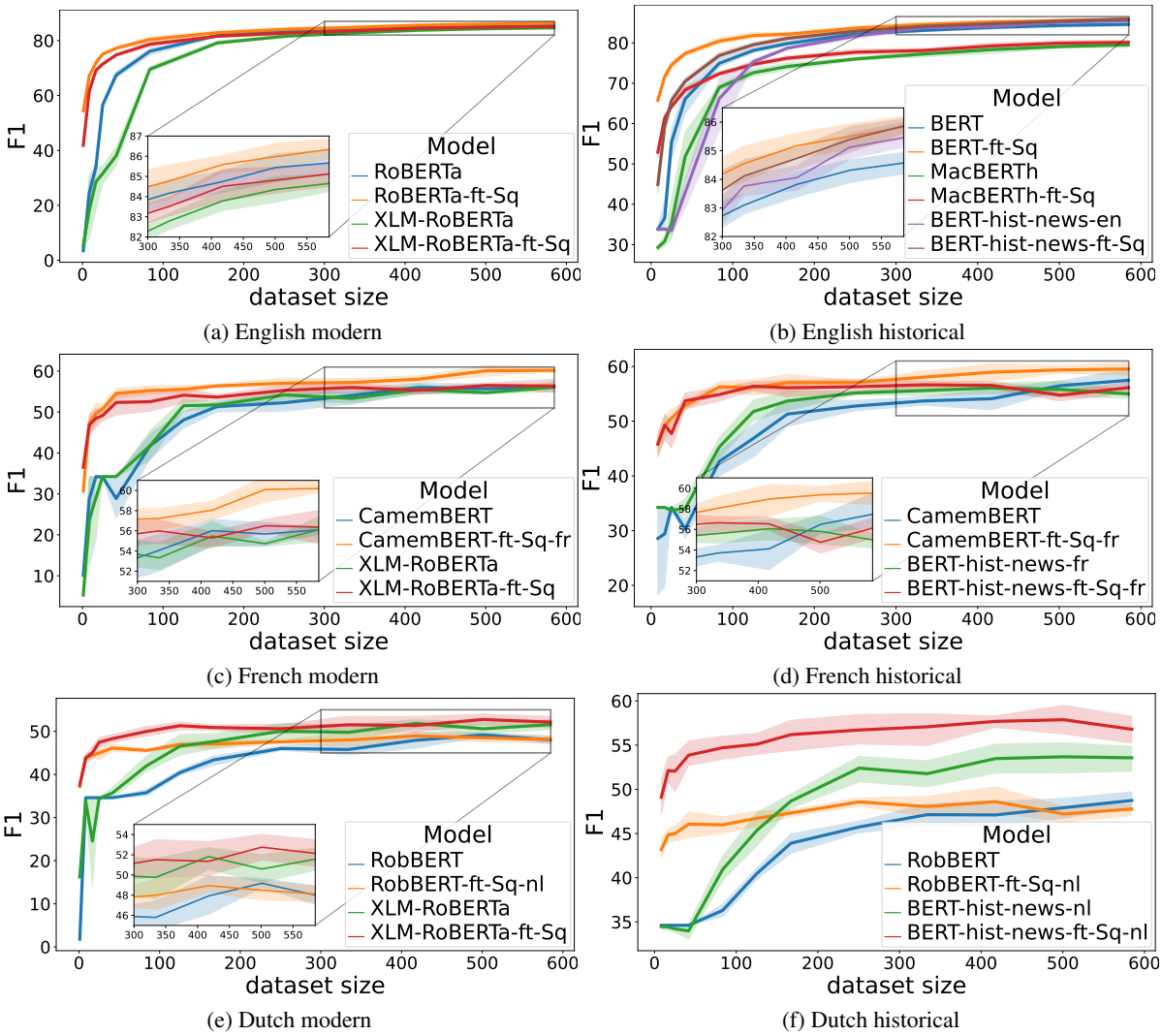
Next, we analyze the results of fine-tuning the models in a fully supervised setting in a single language. Fig. 3a shows the performance of four models on the English evaluation set after being fine-tuned on English training sets of various sizes. All models achieve impressive $F1$ scores even when trained on a small fraction of the training set, further demonstrating the benefit of formulating the task as an extractive QA problem.

Interestingly, the two models intermediately trained on SQuAD perform better than the base models. This trend holds for all dataset sizes but is particularly pronounced in the low-data regime, demonstrating that the SQuAD-based models can generalize with much fewer examples. Comparing Fig. 3a with Tab. 2 further underpins this finding. In addition, we again see that the multilingual models achieve lower $F1$ scores than their monolingual counterparts. Moreover, and unsurprisingly, our results also suggest that the large models perform better than their base versions (Fig. 7 in App. C).

Fig. 3c, 3e repeat some of the trends mentioned above and in §4.1. Again, the models achieve considerably lower $F1$ scores in French and Dutch than in English. While our evaluation of the translation demonstrated the relatively high quality of the process, This gap can still be attributed to noise in the translation process of the train datasets from English to Dutch and French and its bias towards modern language. In addition, for both French and Dutch, the SQuAD-fine-tuned models reach higher $F1$ scores for most (but not all) dataset sizes. Fig. 3e demonstrates, similar to Tab. 2, that multilingual models perform better than the monolingual models for Dutch. Surprisingly, this result cannot be observed in Fig. 3c: A monolingual French model outperforms the two multilingual models by a large margin. Finally, we again see (Fig. 7) that larger language models achieve better results than their smaller versions.

We now investigate language models pre-trained on historical texts and find surprising results (Fig. 3). MacBERTh performs worse than BERT,[7] despite being trained on historical English texts. However, BERT-hist-news-en, trained on historical newspapers, performs better on some data regimes. We further analyze this in §4.5.

[7]For the purpose of fairness, we use BERT rather than RoBERTa for comparison with MacBERTh and BERT-hist-news-en, which are BERT-based models.



(a) English modern

(b) English historical

(c) French modern

(d) French historical

(e) Dutch modern

(f) Dutch historical

Figure 3: Performance of the models in a few-shot setting for the three languages, historical and modern models. All models were trained using their "base" version. "ft-Sq" signifies that the model was fine-tuned on SQuAD or one of its equivalents in French (fr) or Dutch (nl).

The analysis of the French models reveals a slightly different picture (Fig. 3d). However, directly comparing CamemBERT and BERT-hist-news-fr is not possible, as the former is based on RoBERTa while the latter is based on BERT. The results for the Dutch models, presented in Fig. 3f, are particularly intriguing. BERT-hist-news-nl performs significantly better than RobBERT, to the extent that the difference cannot be solely attributed to the differing architectures of the two models.[8] As XLM-RoBERTa also outperforms RobBERT, it seems that this model may not be well-suited for this specific domain. These findings will be further explored in §4.5.

4.3 Semi-Supervised Training

Tab. 3 reveals an interesting result: for English, using the larger unannotated dataset improved the performance of the models for all data sizes. Moreover, tri-training is most effective for English. The picture is less clear, however, for French and Dutch. While using the unannotated data has a positive impact on models trained on the entire dataset, the gains are smaller and tend to be unstable. We leave an in-depth exploration of this for future work.

4.4 Cross-lingual Training

As mentioned in §3.4, we compare two different cross-lingual settings: supervised-only, where we train a cross-lingual model on the English *Runaway Slaves in Britain* dataset while evaluating it on French or Dutch; and MLM settings, where we

[8] RobBERT is based on RoBERTa and BERT-hist-news-nl is based on BERT.

Language	Model	Setting	Dataset size			
			8	16	25	585
en	RoBERTa-base-ft-SQuAD	None	67.13	77.2	80.41	86.33
		Further pre-trained	57.18	76.52	79.93	85.91
		MLM semi-supervised	68.28	78.17	80.8	86.17
		Tri-training	**70.97**	**79.48**	**82.42**	**87.04**
fr	CamemBERT-base-ft-SQuAD	None	**47.3**	**54.55**	55.26	60.19
		Further pre-trained	34.04	49.48	54.04	61.01
		MLM semi-supervised	46.79	48.2	47.11	49.64
		Tri-training	46.76	53.87	**55.98**	**61.58**
	XLM-RoBERTa-base-ft-SQuAD	None	46.8	48.48	49.14	56.36
		Simple cross-lingual	46.08	51.01	51.45	56.28
		MLM cross-lingual	47.0	48.36	48.34	53.98
nl	RobBERT-base-ft-SQuAD	None	44.04	46.12	45.56	48.11
		Further pre-trained	34.61	**46.16**	**48.15**	49.84
		MLM semi-supervised	31.6	41.62	40.22	43.82
	XLM-RoBERTa-base-ft-SQuAD	None	43.73	45.08	47.47	**52.14**
		Simple cross-lingual	43.32	44.84	44.79	46.63
		MLM cross-lingual	**45.94**	45.34	47.1	48.5

Table 3: $F1$ score of the models in semi-supervised and cross-lingual settings. "None" means the model was trained in a standard supervised fashion. For "further pre-trained" we first further train the model on an MLM objective, then train it on our annotated dataset. For "MLM semi-supervised" we train the models on MLM and QA objectives simultaneously, and in "tri-training" we train the models using the tri-training algorithm. This line is missing from the Dutch models as the unlabeled Dutch dataset contains entire newspaper issues and not individual ads. 'Simple cross-lingual" is standard cross-lingual training and "MLM cross-lingual" marks that the model was trained using an MLM-objective in addition to the standard QA loss. Bold marks the best method for a language, while an underline marks the best method for a specific training setting (semi-supervised or cross-lingual). See Tab. 6 and 7 in App. C for evaluation of other models.

also train the model with an MLM-objective using an unlabeled dataset of the target language. Tab. 3 contains the results of this evaluation. Interestingly, it seems that cross-lingual training is more effective when the number of available annotated examples is small. When the entire dataset is being used, however, monolingual training using a translated dataset achieved better performance. Tab. 3 also demonstrates that the MLM settings are preferable over the simple settings in most (but not all) cases.

4.5 Error Analysis

First, we investigate common errors that our most successful models (RoBERTa) make. Fig. 6 in App. C demonstrates that the model struggles with long ads. Perhaps using models that were trained on longer sequences could help with this going forward. A per-attribute analysis, the result of which can be seen in Fig. 4 (pale-colored columns), unsurprisingly suggests that the model finds rare attributes harder to predict (e.g. "ran from region", and compare Fig. 4 to Tab. 8).

Next, we move on to evaluating the generalization capabilities of the models. A per-attribute analysis (Fig. 4, dark-colored columns) reveals that training RoBERTa on SQuAD improved the overall ability of the model to generalize to unseen attributes, probably by utilizing the much broader types of questions that exist in the dataset. However, we also see that the models particularly struggle to generalize to some of them. After closer examination, it seems like these "hard" attributes are either: 1) very rare ("Destination (region)"); 2) non-specific, with possibly more than one span in the ad with the correct type of the answer ("Given name"); or 3) related to topics that are probably not being represented in SQuAD ("Racial descriptor"). We speculate that a more well-tuned conversion of the attributes to natural questions could mitigate some of these issues.

Finally, we compare historical LMs to modern models to understand why MacBERTh underperforms on the *Runaways Slaves in Britain* dataset while BERT-hist-news-en/nl do not. We hypothesize that MacBERTh, trained on a wide range of texts from over 500 years, cannot adapt well to ads written in a language more similar to modern English. Additionally, MacBERTh's training dataset is disproportionately skewed towards texts from

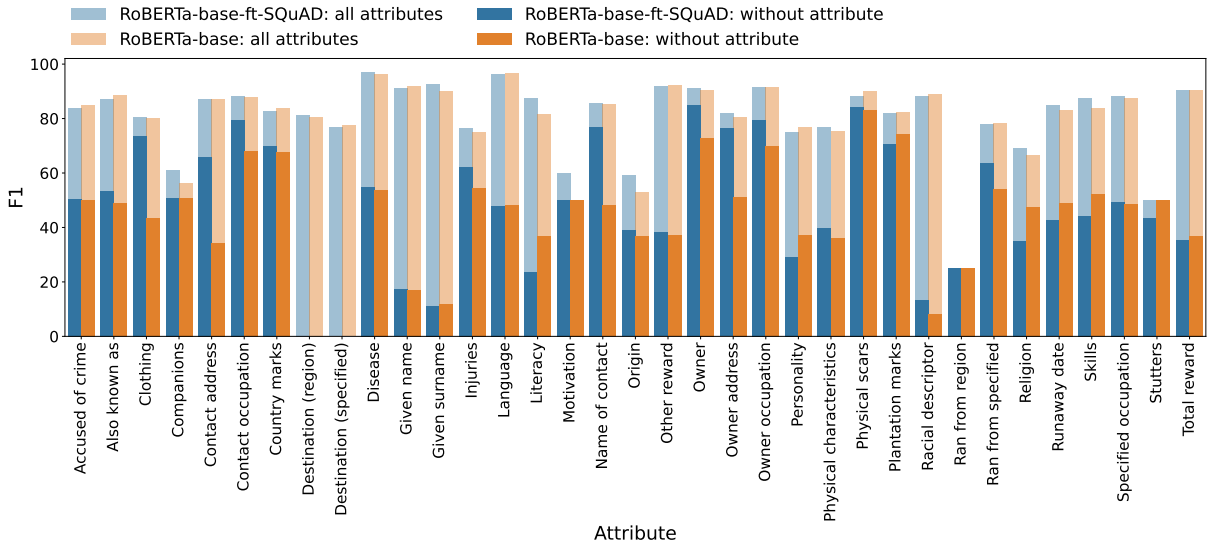


Figure 4: The generalization capabilities of RoBERTa in a fully-supervised setting. The columns in pale color describe the performance of the models on the attribute with standard training, whereas the columns in darker color describe the performance on the attribute of a model that was not trained on the attribute (generalization).

1600-1690 and 1830-1950, while texts from 1700-1850 (the period corresponding to our dataset) are scarce. In contrast, BERT-hist-news-en/nl were trained on datasets containing mostly 19th-century newspapers, a domain and period closer to our.

To validate this, we calculate the perplexity of our dataset w.r.t. the models (technical details in App. A.1). Indeed, the perplexity of our English newspaper ads dataset w.r.t. MacBERTh is higher (16.47) than the perplexity w.r.t. BERT (15.32) and BERT-hist-news-en (5.65). A similar picture emerges for Dutch: the perplexity of our Dutch test dataset of newspaper ads w.r.t RobBERT was significantly higher (49.53) than the perplexity w.r.t. BERT-hist-news-nl (5.12).

5 Conclusions

In this work, we address the unique challenges of event extraction from historical texts in different languages. We start by developing a new multilingual dataset in English, French, and Dutch of events, consisting of newspaper adverts reporting on enslaved people escaping their enslavers. We then demonstrate the benefits of framing the problem as an extractive QA task. We show that even with scarcely annotated data, this formulation can achieve surprisingly good results by leveraging existing datasets and models for modern languages. Finally, we show that cross-lingual low-resource learning for historical languages is highly challenging, and machine translation of the historical datasets to the considered target languages is, in practice, often the best-performing solution.

Limitations

We see four main limitations regarding our work. First, we have evaluated our models on a dataset containing events of one type only. It remains to be seen how applicable our formulation and methods are to other historical datasets and event types. Second, given the nature of the historical question our dataset targets, it contains documents only from one language family. Extending our methodology to languages from other language families might pose further challenges in terms of multilinguality. Third, our method relies heavily on automatic translation tools, which are biased toward translating historical texts into modern language. This can negatively affect the performance of our models. Lastly, in real-life cases, machine readable historical texts are often extremely noisy, suffering from high level of OCR errors and other text extraction mistakes. Conversely, we have tested our methods on relatively clean datasets, with the unannotated Dutch material as the only exception. We leave a more thorough study on how well our proposed methods are suitable for noisy text to future work.

Ethical Considerations

Studying texts about the history of slavery poses ethical issues to historians and computer scientists alike since people of color still suffer consequences

of this history in the present, not least because of lingering racist language (Alim et al., 2016, 2020).

As researchers, we know that an important ethical task is to develop sound NLP tools that can aid in the examination of historical texts containing racist language, while endeavoring at all costs not to reproduce or perpetuate such racist language through the very tools we develop.

The enslaved people described in the newspapers adverts used in this study were alive centuries ago, so any immediate issues related to their privacy and personal data protection do not apply. Nonetheless, the newspaper adverts studied here were posted by the oppressors of the people who tried to liberate themselves, and contain many examples of highly racist and demeaning language.

Acknowledgements

This work is partly funded by the Danish National Research Foundation (DNRF 138). Isabelle Augenstein is further supported by the Pioneer Centre for AI, DNRF grant number P1.

References

H. Samy Alim, Angela Reyes, and Paul V. Kroskrity, editors. 2020. *The Oxford Handbook of Language and Race*. Oxford University Press. Publication Title: The Oxford Handbook of Language and Race.

H. Samy Alim, John R. Rickford, and Arnetha F. Ball, editors. 2016. *Raciolinguistics: How Language Shapes Our Ideas About Race*. Oxford University Press, New York.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *CoRR*, abs/1910.11856.

Blouin Baptiste, Benoit Favre, Jeremy Auguste, and Christian Henriot. 2021. Transferring modern named entity recognition to the historical domain: How to take the step? In *Workshop on Natural Language Processing for Digital Humanities (NLP4DH)*.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Marcel Bollmann. 2019. A large-scale comparison of historical text normalization systems. 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 3885–3898, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcel Bollmann and Anders Søgaard. 2016. Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 131–139, Osaka, Japan. The COLING 2016 Organizing Committee.

Marcel Bollmann, Anders Søgaard, and Joachim Bingel. 2018. Multi-task learning for historical text normalization: Size matters. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 19–24.

Nadav Borenstein, Karolina Stańczak, Thea Rolskov, Natacha Klein Käfer, Natalia da Silva Perez, and Isabelle Augenstein. 2023. Measuring intersectional biases in historical documents. *Association for Computational Linguistics*.

Emanuela Boros, Nhu Khoa Nguyen, Gaël Lejeune, and Antoine Doucet. 2022. Assessing the impact of ocr noise on multilingual event detection over digitised documents. *International Journal on Digital Libraries*, pages 1–26.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Agata Cybulska and Piek Vossen. 2011. Historical event extraction from text. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 39–43.

Mark Darling, Marieke Meelen, and David Willis. 2022. Towards coreference resolution for early irish. In *Proceedings of the LREC Conference*. LREC Conference.

Francesco De Toni, Christopher Akiki, Javier De La Rosa, Clémentine Fourrier, Enrique Manjavacas, Stefan Schweter, and Daniel Van Strien. 2022. Entities, dates, and languages: Zero-shot on historical texts with t0. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 75–83, virtual+Dublin. Association for Computational Linguistics.

Deepset. 2022a. Multilingual xlm-roberta base for qa on various languages. `https://huggingface.co/deepset/xlm-roberta-base-squad2`. Accessed: 2022-06-01.

Deepset. 2022b. Roberta base for qa. https://huggingface.co/deepset/roberta-base-squad2. Accessed: 2022-06-01.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. 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.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. 2020. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online. Association for Computational Linguistics.

Senka Drobac, Pekka Kauppinen, and Krister Lindén. 2017. Ocr and post-correction of historical finnish texts. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 70–76.

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Xinya Du, Alexander Rush, and Claire Cardie. 2021. GRIT: Generative role-filler transformers for document-level event entity extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 634–644, Online. Association for Computational Linguistics.

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2021. Named entity recognition and classification on historical documents: A survey. *arXiv preprint arXiv:2109.11406*.

Etalab. 2021. Piaf - le dataset francophone de questions-réponses. Accessed: 2022-12-10.

Anne Gerritsen. 2012. Scales of a local: the place of locality in a globalizing world. *A Companion to World History*, pages 213–226.

Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar. 2021. Lemmatization of historical old literary Finnish texts in modern orthography. In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 189–198, Lille, France. ATALA.

Christian Hardmeier. 2016. A neural model for part-of-speech tagging in historical texts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 922–931, Osaka, Japan. The COLING 2016 Organizing Committee.

Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, and Franciska De Jong. 2011. An overview of event extraction from text. *DeRiVE@ ISWC*, pages 48–57.

Ali Kabbadj. 2021. French-squad : French machine reading for question answering.

Markus Krug, Frank Puppe, Fotis Jannidis, Luisa Macharowsky, Isabella Reger, and Lukas Weimar. 2015. Rule-based coreference resolution in german historic novels. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 98–104.

Viet Lai, Minh Van Nguyen, Heidi Kaufman, and Thien Huu Nguyen. 2021. Event extraction from historical texts: A new dataset for black rebellions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2390–2400.

Julia Laite. 2020. The emmet's inch: Small history in a digital age. *Journal of Social History*, 53(4):963–989.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838, Online. Association for Computational Linguistics.

Ying Lin, Heng Ji, Fei Huang, and Lingfei Wu. 2020. A joint neural model for information extraction with global features. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7999–8009, Online. Association for Computational Linguistics.

Jian Liu, Yubo Chen, Kang Liu, Wei Bi, and Xiaojiang Liu. 2020. Event extraction as machine reading comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1641–1651, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lijun Lyu, Maria Koutraki, Martin Krickl, and Besnik Fetahu. 2021. Neural ocr post-hoc correction of historical corpora. *Transactions of the Association for Computational Linguistics*, 9:479–493.

Enrique Manjavacas and Lauren Fonteyn. 2022. Adapting vs. Pre-training Language Models for Historical Languages. *Journal of Data Mining & Digital Humanities*, NLP4DH.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. Camembert: a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Simon P. Newman, Stephen Mullen, Nelson Mundell, and Roslyn Chapman. 2019. Runaway Slaves in Britain: bondage, freedom and race in the eighteenth century. https://www.runaways.gla.ac.uk. Accessed: 2022-12-10.

Thien Nguyen and Ralph Grishman. 2018. Graph convolutional networks with argument-aware pooling for event detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. *arXiv e-prints*, page arXiv:1606.05250.

Ricardo Rei, Ana C Farinha, Chrysoula Zerva, Daan van Stigt, Craig Stewart, Pedro Ramos, Taisiya Glushkova, André F. T. Martins, and Alon Lavie. 2021. Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1030–1040, Online. Association for Computational Linguistics.

Alexander Robertson and Sharon Goldwater. 2018. Evaluating historical text normalization systems: How well do they generalize? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 720–725, New Orleans, Louisiana. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. 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 3118–3135, Online. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization.

Stefan Schweter. 2020. Europeana bert and electra models. https://huggingface.co/dbmdz/bert-base-french-europeana-cased. Accessed: 2022-12-10.

Stefan Schweter. 2021a. Language model for historic dutch. https://huggingface.co/dbmdz/bert-base-historic-dutch-cased. Accessed: 2022-12-10.

Stefan Schweter. 2021b. Language model for historic english. https://huggingface.co/dbmdz/bert-base-historic-english-cased. Accessed: 2022-12-10.

Roxane Segers, Marieke Van Erp, Lourens Van Der Meij, Lora Aroyo, Jacco van Ossenbruggen, Guus Schreiber, Bob Wielinga, Johan Oomen, and Geertje Jacobs. 2011. Hacking history via event extraction. In *Proceedings of the sixth international conference on Knowledge capture*, pages 161–162.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634, Florence, Italy. Association for Computational Linguistics.

Rachele Sprugnoli and Sara Tonelli. 2019. Novel Event Detection and Classification for Historical Texts. *Computational Linguistics*, 45(2):229–265.

Yixuan Tang, Hwee Tou Ng, and Anthony Tung. 2021. Do multi-hop question answering systems know how to answer the single-hop sub-questions? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3244–3249, Online. Association for Computational Linguistics.

Jianshu Weng and Bu-Sung Lee. 2011. Event detection in twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 401–408.

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.

Wei Xiang and Bang Wang. 2019. A survey of event extraction from text. *IEEE Access*, 7:173111–173137.

Yi Yang and Jacob Eisenstein. 2016. Part-of-speech tagging for historical english. In *HLT-NAACL*, pages 1318–1328.

Zhisong Zhang, Xiang Kong, Zhengzhong Liu, Xuezhe Ma, and Eduard Hovy. 2020. A two-step approach for implicit event argument detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7479–7485, Online. Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics.

Zhi-Hua Zhou and Ming Li. 2005. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Transactions on knowledge and Data Engineering*, 17(11):1529–1541.

Appendix

A Reproducibility

A.1 Calculating Perplexity

To calculate the (pseudo)-perplexity of a sentence $S = w_1 w_2 w_3 ... w_n$ w.r.t. a masked language model, we used the following formula

$$PP_{(S)} = \left(\prod_{i=1}^{n} P(w_i | S_{-i}) \right)^{-1/n} = \left(\prod_{i=1}^{n} P_{\text{MLM}}(w_i | S_{-i}) \right)^{-1/n} \quad (1)$$

where S_{-i} is the sentence S masked at token i. To calculate the perplexity of an entire corpus $C = S^1, S^2, ..., S^m$ w.r.t. a masked language model we notice that $P(w_i^j | C_{-(j,i)}) = P(w_i^j | S_{-i}^j)$, where $C_{-(j,i)}$ is the corpus C with sentence j masked at location i.

Therefore,

$$PP_{(C)} = \left(\prod_{j=1}^{m} \prod_{i=1}^{|S^j|} P_{\text{MLM}}(w_i^j | S_{-i}^j) \right)^{-1/k} \quad (2)$$

where k is the total number of tokens in the corpus, i.e. $k = \sum_{j=1}^{m} |S^j|$.

Notice, that in the log space this formula becomes equivalent to the average of the negative log likelihoods:

$$\log\left(PP_{(C)}\right) = \frac{1}{k} \left(\sum_{j=1}^{m} \sum_{i=1}^{|S^j|} \text{NLL}_{\text{MLM}}(w_i^j | S_{-i}^j) \right) \quad (3)$$

where NLL_{MLM} is the negative log likelihood, which in many cases equal to passing the output of the language model to a standard cross entropy loss.

A.2 Translation of the Annotated Dataset

A.2.1 Translation Process

Each sample in the annotated Runaways dataset follows the SQuAD-v2 scheme, and contains a context c (the ad's text), a question q (one of the attributes) and an answer a such that a appears in c (a might also be the empty string). We used the publicly available Google Translate API[9] to translate the samples into the target languages. We also considered using Facebook's NLLB model (Costa-jussà et al., 2022),[10] but it performed noticeably worse. See below for more details regarding evaluating the quality of the translation.

Language	Translation tool	COMET score
French	Google Translate	**0.014**
	NLLB	0.01
Dutch	Google Translate	**0.017**
	NLLB	0.01

Table 4: Evaluation of the translation quality using COMET (higher is better).

Unfortunately, simply translating (c, q, a) from English to the target language is not enough. In some cases, translation of the context and the answer are not always aligned. That is, translating c to c^t and a to a^t results in a pair for which a^t does not appear verbatim in c^t. In those cases we try to find a span of text $\hat{a}^t$ in c^t such that $\hat{a}^t$ is similar to a^t (and therefore, hopefully the correct answer to the question q).

To achieve this, we use fuzzy string matching[11] to find $\hat{a}^t$. Specifically, we did the following. First, we calculated $k = \max(|a^t|, |a|)$, and extracted all the k-grams from c^t. Then, we used fuzzy string search to find the k-gram that is most similar to a^t, with a score of at least 0.5. We then assign $k = k + 1$ and repeat the process five times, finally returning the match with the highest score. If no match was found, we assign $a^t = a$ (this is useful in cases where the answer is a name, a date etc.) and repeat the above-mentioned algorithm. If again no match is found the matching has failed and we discard the sample.

Finally, we opted to manually translate q as the number of different questions in our dataset is relatively low.

A.2.2 Evaluation of the Translation

We evaluated several translation tools. Based on preliminary evaluation, we determined that Google Translate and Facebook's NLLB model were the most promising options, as other methods either did not meet the minimum desired quality or were

[9] using the deep-translator package, https://deep-translator.readthedocs.io/en/latest/

[10] we used the 3.3b parameters variant https://huggingface.co/facebook/nllb-200-3.3B, as it was the biggest model available we could load on our machine

[11] using https://pypi.org/project/fuzzywuzzy/

Language	Translation tool	Accuracy	Fluency
French	Google Translate	**4.5**	**3.4**
	NLLB	3.7	**3.4**
Dutch	Google Translate	**4.8**	**4.2**
	NLLB	3.5	3.3

Table 5: Evaluation of the translation quality using human raters (higher is better).

difficult to run on large datasets. We evaluated the two translation schemes using automatic tools and human raters. Both metrics demonstrated the superiority of Google Translate over NLLB in terms of accuracy and fluency, as shown below.

Automatic method We used COMET, a state-of-the-art reference-free automatic translation evaluation tool (Rei et al., 2021), and used it to evaluate the quality of translating the original English ads to French and Dutch. Tab. 4 contains the result of running the model, demonstrating the higher quality of the translations produced by Google Translate compared to NLLB.

Human evaluation We asked native speakers to rate 20 translations of ads on a scale of 1-5 for accuracy and fluency. They were instructed to give a translation a fluency score of 5 if it is as fluent as the original English text, and 1 if it was barely readable. Similarly, they were instructed to give an accuracy score of 5 if all the ad's attributes describing the self-liberation event were translated correctly and 1 if almost none of them were. Tab. 5 demonstrate not only that Google Translate is the better translation tool, but also that the accuracy and fluency of the tool are objectively good.

A.3 Zero-Shot Inference with T0++

T0++ is a prompt-based encoder-decoder LM developed as part of the BigScience project (Sanh et al., 2021). One of the tasks that T0++ was trained on is extractive QA. To train the model on an extractive QA task, the designers of T0++ converted an extractive QA dataset, such as SQuAD into a prompt format. Each example with question q, context c and answer a in the dataset was placed into one of several possible templates, such as "*Given the following passage: {c}, answer the following question. Note that the answer is present within the text. Question: {q}*". T0++ was trained to generate a given the template as a prompt.

To perform inference with T0++ with our datasets we followed De Toni et al. (2022) and the original training routine of T0++. We converted the dataset to prompts using one of the templates that were used to train the model on extractive QA, and tried to map T0++'s prediction into the original context. As De Toni et al. (2022) we tried two mapping methods – an exact matching, where we consider T0++'s prediction valid only if the prediction appears verbatim in the context; and a fuzzy matching method, where some variation is allowed. If no match is found we discard the prediction and assume that the answer to the question does not exist in the context. In Tab. 2 we report the result of the "exact match" method, which performed better in practice.

A.4 Training Details

We specify here the hyper-parameters that were used to train our models for reproduciblity purpose.

- Number of epochs: 5
- Learning rate: $5e-5$
- Batch size: 32 (for models trained with an additional MLM objective: 16 for each objective)
- Weight decay: 0
- Sequence length: 256

Other settings were set to their default values (when using Huggingface's Trainer[12] object).

B Annotation Guidelines

Here we describe the annotation guidelines that were used for creating the evaluation set of the multilingual dataset. The experts were instructed to follow the same annotation scheme that was used to create the *Runaway slaves in Britain* dataset. That is, given an ad, they were asked to find and mark in the ad the same 50 attributes that exist in the Runaway dataset (App. D). More specifically, we asked the experts to familiarize themselves with the 50 attributes and ensured they understood them. We also supplied them with an English example to demonstrate how to perform the task and asked them to annotate the other ads in their respective language. To add an attribute, the annotators had to mark a span of text with their mouse and click on an attribute name from a color-coded list. Each attribute can be annotated more than once in each ad. Fig. 5 shows a screenshot of the annotation tool that we used (Markup[13]) and the English example.

[12] https://huggingface.co/docs/transformers/main_classes/trainer

[13] https://getmarkup.com/

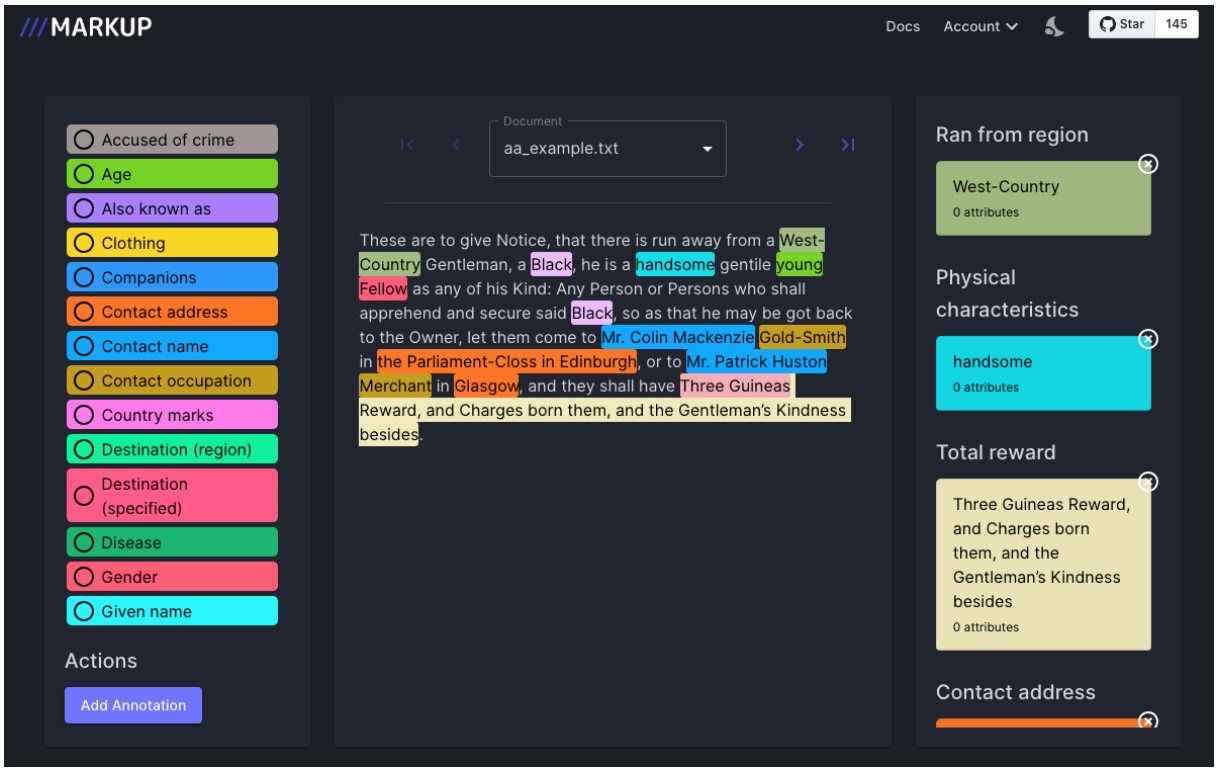


Figure 5: A screenshot of the annotation tool used by the experts. The ad shown here is an example that was presented to each expert, and they were instructed to annotate the other ads similarly.

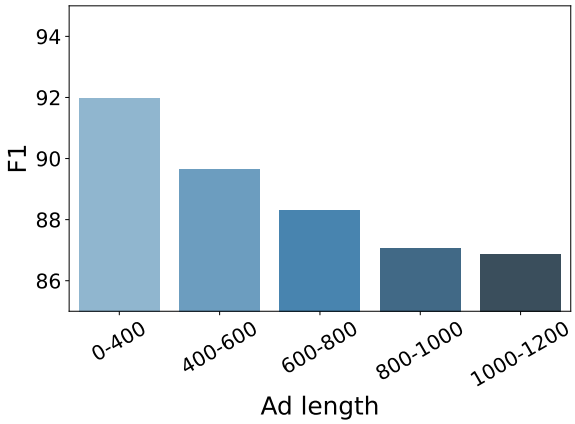


Figure 6: Performance of RoBERTa, fine-tuned on SQuAD-v2, on the English dataset. The longer the ad, the worse the model performs.

C Additional Results

D Attributes

Tab. 8 lists the different attributes that we wish to extract from the advertisements. The column "Question" describes the question that we feed the models in order to retrieve that attribute, and #Annotated contains the number of occurrences of the attribute in the annotated dataset.

Language	Model	Setting	Dataset size			
			8	16	25	585
en	RoBERTa-base	None	24.42	67.43	76.1	85.66
		further pre-trained	15.22	69.52	77.59	85.85
		MLM	33.13	71.32	78.06	86.22
		tri-training	37.27	73.72	79.65	86.1
	RoBERTa-base-ft-SQuAD2	None	67.13	77.2	80.41	86.33
		further pre-trained	57.18	76.52	79.93	85.91
		MLM	68.28	78.17	80.8	86.17
		tri-training	**70.97**	**79.48**	**82.42**	**87.04**
fr	CamemBERT-base	None	28.75	28.87	41.68	56.1
		further pre-trained	26.33	24.13	40.82	57.93
		MLM	23.38	34.24	44.13	58.5
		tri-training	17.11	30.9	48.77	56.98
	CamemBERT-base-ft-SQuAD2	None	**47.3**	**54.55**	55.26	60.19
		further pre-trained	34.04	49.48	54.04	61.01
		MLM	46.79	48.2	47.11	49.64
		tri-training	46.76	53.87	**55.98**	**61.58**
nl	RobBERT-base	None	34.61	34.61	35.76	48
		further pre-trained	34.61	34.24	37.03	49.02
		MLM	42.84	43.29	43.67	46.35
	RobBERT-base-ft-SQuAD2	None	**44.04**	46.12	45.56	48.11
		further pre-trained	34.61	**46.16**	**48.15**	**49.84**
		MLM	31.6	41.62	40.22	43.82

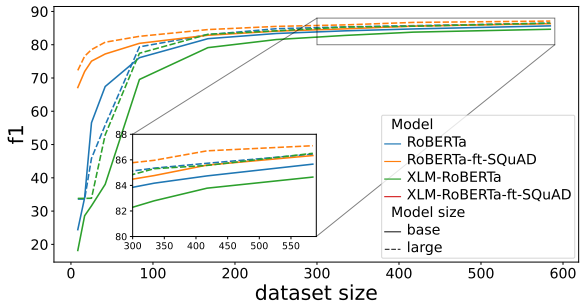
Table 6: $F1$ score of the models in semi-supervised settings. "None" means that no unannotated data were used. In "further pre-trained" we first further pre-train the model on an MLM objective and then fine-tune it on our annotated dataset. In "MLM" we train the models on an MLM and QA objective simultaneously. Finally, in "tri-training" we train the models using the tri-training algorithm. This line is missing from the Dutch models as the unlabeled Dutch dataset contains entire newspaper issues and not individual ads

Language	Model	Setting	Dataset size			
			8	16	25	585
fr	CamemBERT-base	None	28.75	34.24	34.13	56.1
	CamemBERT-base-ft-SQuAD-fr	None	**47.3**	49.68	50.8	**60.2**
	XLM-RoBERTa-base	None	23.63	29.06	34.24	56.1
		Simple	22.17	23.98	29.19	54.73
		MLM	33.36	29.93	25.57	55.63
	XLM-RoBERTa-base-ft-SQuAD-fr	None	46.8	48.48	49.14	56.36
		Simple	46.08	**51.01**	**51.45**	56.28
		MLM	47.0	48.36	48.34	53.98
nl	RobBERT-base	None	34.62	34.62	34.62	48.0
	RobBERT-base-ft-SQuAD-nl	None	44.05	44.4	45.0	48.11
	XLM-RobBERT-base	None	33.8	24.55	34.42	51.56
		Simple	17.23	26.3	33.15	44.45
		MLM	37.66	45.21	45.76	46.31
	XLM-RobBERT-base-ft-SQuAD-nl	None	43.73	45.08	**47.47**	**52.14**
		Simple	43.32	44.84	44.79	46.63
		MLM	**45.94**	**45.34**	47.1	48.5

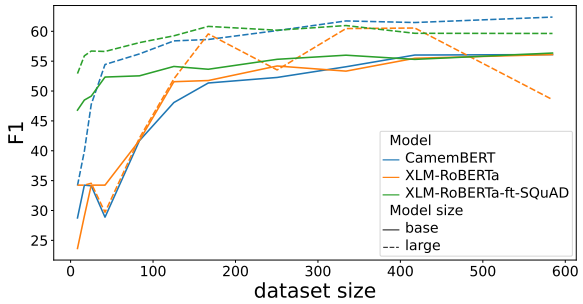
Table 7: $F1$ score of the models in different cross-lingual settings. "None" means that no cross-lingual training were used. "Simple" is standard cross-lingual training and "MLM" marks that the model was trained using an MLM-objective in addition to the standard QA loss.

Attribute	Question	#Annotated
Accused of crime	What crimes did the person commit?	107
Also known as	What other aliases does the person have?	103
Clothing	What clothes did the person wear?	656
Companions	What are the names of the person's friends?	49
Contact address	Where does the contact person of the ad live?	740
Contact occupation	What does the contact of the ad do for a living?	278
Country marks	What country marks does the person have?	63
Destination (region)	What is the destination region of the person?	15
Destination (specified)	What is the name of the destination?	118
Disease	What kind of diseases does the person have?	91
Given name	What is the given name of the person?	693
Given surname	What is the last name of the person?	196
Injuries	How was the person injured?	63
Language	What are the communication skills of the person?	319
Literacy	What is the literacy level of the person?	8
Motivation	Why did the person escape his owner?	4
Name of contact	Who is the contact person for the ad?	678
Origin	Where does the person originate from?	28
Other reward	What other rewards were offered?	382
Owner	Who is the owner of the person?	395
Owner address	Where does the owner of the person live?	270
Owner occupation	What does the owner of the person do for a living?	78
Personality	What are the personality traits of the person?	15
Physical characteristics	What are the physical characteristics of the person?	568
Physical scars	What scars does the person have?	131
Plantation marks	What plantation marks does the person have?	23
Racial descriptor	What is the ethnicity of the person?	807
Ran from region	What is the name of the region the person escaped from?	3
Ran from specified	What is the name of the place the person escaped from?	406
Religion	What is the religion of the person?	13
Runaway date	What was the date of the event?	15
Skills	What is the set of skills of the person?	55
Specified occupation	What does the person do for a living?	98
Stutters	Does the person stutter?	22
Total reward	How much reward is offered?	780

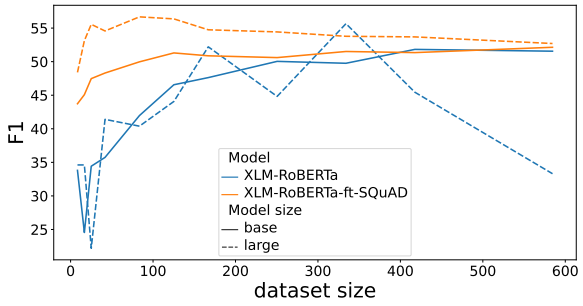
Table 8: The attributes of the *Runaways* dataset



(a) English



(b) French



(c) Dutch

Figure 7: Performance of models of different sizes on the Runaway dataset. The large models perform better than the base models for almost all cases in English, but tend to be more unstable in the other two languages. Unfortunately, not every model in French and Dutch is available in its larger version. Figures 7b and 7c include only the models for which both the base and the large version exist.

ACL 2023 Responsible NLP Checklist

A For every submission:

☐ A1. Did you describe the limitations of your work?
Left blank.

☐ A2. Did you discuss any potential risks of your work?
Left blank.

☐ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

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

B ☐ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
Left blank.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Left blank.

☐ 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)?
Left blank.

☐ 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?
Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Left blank.

☐ 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.
Left blank.

C ☐ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

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?
Left blank.

☐ 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?
Left blank.

☐ 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.)?
Left blank.

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

Left blank.

☐ 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.?
Left blank.

☐ 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)?
Left blank.

☐ 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?
Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Left blank.