Multilingual Event Extraction from Historical Newspaper Adverts

Borenstein, Nadav; Da Silva Perez, Natália; Augenstein, Isabelle

Published in:
Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics

DOI:
10.18653/v1/2023.acl-long.574

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
https://doi.org/10.18653/v1/2023.acl-long.574

Download date: 19. apr., 2024
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Analyzing large corpora of historical documents can provide invaluable insights on past events in multiple resolutions, from the life of an individual to processes on a global scale (Borenstein et al., 2023;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-
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 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
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; 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...
to match the format of SQuAD-v2 (Rajpurkar et al., 2016), a large benchmark for extractive QA. 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, 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, 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Language</th>
<th>#Labeled ads</th>
<th>#Labeled attributes</th>
<th>#Unlabeled ads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Runaways Slaves in Britain</td>
<td>en</td>
<td>835</td>
<td>8270</td>
<td>0</td>
</tr>
<tr>
<td>Runaways Slaves in Britain</td>
<td>fr (translated)</td>
<td>834</td>
<td>8238</td>
<td>0</td>
</tr>
<tr>
<td>Runaways Slaves in Britain</td>
<td>nl (translated)</td>
<td>834</td>
<td>8234</td>
<td>0</td>
</tr>
<tr>
<td>Marronage</td>
<td>en</td>
<td>0</td>
<td>0</td>
<td>3026</td>
</tr>
<tr>
<td>Marronage</td>
<td>fr</td>
<td>41</td>
<td>313</td>
<td>19066</td>
</tr>
<tr>
<td>Delpher</td>
<td>nl</td>
<td>44</td>
<td>272</td>
<td>2742 issues</td>
</tr>
</tbody>
</table>

Table 1: Sizes of the different datasets.
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. 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.

---

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

---

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.

---

Table 2: Zero-shot performance of different models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Fine-tune data</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OneIE</td>
<td>N/A</td>
<td>51.90</td>
</tr>
<tr>
<td>T0++</td>
<td>N/A</td>
<td>33.69</td>
</tr>
<tr>
<td>RoBERTa-base</td>
<td>SQuAD-v2</td>
<td>54.35</td>
</tr>
<tr>
<td>RoBERTa-large</td>
<td>SQuAD-v2</td>
<td>56.42</td>
</tr>
<tr>
<td>XLM-RoBERTa-base</td>
<td>SQuAD-v2</td>
<td>41.84</td>
</tr>
<tr>
<td>XLM-RoBERTa-large</td>
<td>SQuAD-v2</td>
<td>55.10</td>
</tr>
<tr>
<td>fr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T0++</td>
<td>N/A</td>
<td>32.26</td>
</tr>
<tr>
<td>CamemBERT-base</td>
<td>PIAF-v1.1</td>
<td>30.65</td>
</tr>
<tr>
<td></td>
<td>FQuAD-v1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SQuAD-FR</td>
<td></td>
</tr>
<tr>
<td>XLM-RoBERTa-base</td>
<td>SQuAD-v2</td>
<td>36.51</td>
</tr>
<tr>
<td>XLM-RoBERTa-large</td>
<td>SQuAD-v2</td>
<td>44.52</td>
</tr>
<tr>
<td>nl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T0++</td>
<td>N/A</td>
<td>29.28</td>
</tr>
<tr>
<td>RobBERT-base</td>
<td>SQuAD-NL</td>
<td>37.21</td>
</tr>
<tr>
<td>XLM-RoBERTa-base</td>
<td>SQuAD-v2</td>
<td>37.56</td>
</tr>
<tr>
<td>XLM-RoBERTa-large</td>
<td>SQuAD-v2</td>
<td>40.42</td>
</tr>
</tbody>
</table>

---
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.\footnote{RobBERT is based on RoBERTa and BERT-hist-news-nl is based on BERT.} 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 \S 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 \S 3.4, we compare two different cross-lingual settings: supervised-only, where we train a cross-lingual model on the English Run-away Slaves in Britain dataset while evaluating it on French or Dutch; and MLM settings, where we...
<table>
<thead>
<tr>
<th>Language</th>
<th>Model</th>
<th>Setting</th>
<th>Dataset size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>8 16 25 585</td>
</tr>
<tr>
<td>en</td>
<td>RoBERTa-base-ft-SQuAD</td>
<td>None</td>
<td>67.13 77.2 80.41 86.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Further pre-trained</td>
<td>57.18 76.52 79.93 85.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MLM semi-supervised</td>
<td>68.28 78.17 80.8 86.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tri-training</td>
<td><strong>70.97 79.48 82.42 87.04</strong></td>
</tr>
<tr>
<td>fr</td>
<td>CamemBERT-base-ft-SQuAD</td>
<td>None</td>
<td><strong>47.3 54.55</strong> 55.26 60.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Further pre-trained</td>
<td>34.04 49.48 54.04 61.01</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MLM semi-supervised</td>
<td>46.79 48.2 47.11 49.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tri-training</td>
<td>46.76 53.87 <strong>55.98 61.58</strong></td>
</tr>
<tr>
<td>nl</td>
<td>RobBERT-base-ft-SQuAD</td>
<td>None</td>
<td>44.04 46.12 45.56 48.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Further pre-trained</td>
<td>34.61 <strong>46.16 48.15</strong> 49.84</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MLM semi-supervised</td>
<td>31.6 41.62 40.22 43.82</td>
</tr>
<tr>
<td>nl</td>
<td>XLM-RoBERTa-base-ft-SQuAD</td>
<td>Simple cross-lingual</td>
<td>43.73 45.08 47.47 <strong>52.14</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MLM cross-lingual</td>
<td>43.32 44.84 44.79 46.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td><strong>45.94 45.34</strong> 47.1 48.5</td>
</tr>
</tbody>
</table>

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.

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...
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 sig-
nificantly 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 multi-
lingual 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 prob-
lem as an extractive QA task. We show that even
with scarcely annotated data, this formulation can
achieve surprisingly good results by leveraging ex-
stisting datasets and models for modern languages.
Finally, we show that cross-lingual low-resource
learning for historical languages is highly chal-
lenging, 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. Sec-
ond, 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 multilingual-
ity. Third, our method relies heavily on automatic
translation tools, which are biased toward translat-
ing historical texts into modern language. This can
negatively affect the performance of our models.
Lastly, in real-life cases, machine readable histori-
ical 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


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

\[
PP(S) = \left( \prod_{i=1}^{n} P(w_i|S_{-i}) \right)^{-1/n} = \left( \prod_{i=1}^{n} P_{\text{MLM}}(w_i|S_{-i}) \right)^{-1/n} \tag{1}
\]

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 \( P(w_i|C_{-(j,i)}) = P(w_i|S^j_{-i}) \), where \( C_{-(j,i)} \) is the corpus \( C \) with sentence \( j \) masked at location \( i \).

Therefore,

\[
PP(C) = \left( \prod_{j=1}^{m} \prod_{i=1}^{n} P_{\text{MLM}}(w_i^j|S^j_{-i}) \right)^{-1/k} \tag{2}
\]

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:

\[
\log(PP(C)) = \frac{1}{k} \left( \sum_{j=1}^{m} \sum_{i=1}^{n} \text{NLL}_{\text{MLM}}(w_i^j|S^j_{-i}) \right) \tag{3}
\]

where \( \text{NLL}_{\text{MLM}} \) 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.

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 \( c' \) and \( a \) to \( a' \) results in a pair for which \( a' \) does not appear verbatim in \( c' \). In those cases we try to find a span of text \( \hat{a}' \) in \( c' \) such that \( \hat{a}' \) is similar to \( a' \) (and therefore, hopefully the correct answer to the question \( q \)).

To achieve this, we use fuzzy string matching\(^11\) to find \( \hat{a}' \). Specifically, we did the following. First, we calculated \( k = \max(|a'|, |a|) \), and extracted all the k-grams from \( c' \). Then, we used fuzzy string search to find the k-gram that is most similar to \( a' \), 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 \( a' = a \) (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

<table>
<thead>
<tr>
<th>Language</th>
<th>Translation tool</th>
<th>COMET score</th>
</tr>
</thead>
<tbody>
<tr>
<td>French</td>
<td>Google Translate</td>
<td>0.014</td>
</tr>
<tr>
<td></td>
<td>NLLB</td>
<td>0.01</td>
</tr>
<tr>
<td>Dutch</td>
<td>Google Translate</td>
<td>0.017</td>
</tr>
<tr>
<td></td>
<td>NLLB</td>
<td>0.01</td>
</tr>
</tbody>
</table>

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

\(^10\)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/
The translation quality was evaluated using human raters (higher is better).

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 reproducibility purpose.

- Number of epochs: 5
- Learning rate: \( 5 \times 10^{-5} \)
- 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\textsuperscript{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\textsuperscript{13}).

\textsuperscript{12}https://huggingface.co/docs/transformers/main_classes/trainer
\textsuperscript{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.
<table>
<thead>
<tr>
<th>Language</th>
<th>Model</th>
<th>Setting</th>
<th>Dataset size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>en</td>
<td>RoBERTa-base</td>
<td>None</td>
<td>24.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td>further pre-trained</td>
<td>15.22</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MLM</td>
<td>33.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tri-training</td>
<td>37.27</td>
</tr>
<tr>
<td>en</td>
<td>RoBERTa-base-ft-SQuAD2</td>
<td>None</td>
<td>67.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>further pre-trained</td>
<td>57.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MLM</td>
<td>68.28</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tri-training</td>
<td>70.97</td>
</tr>
<tr>
<td>fr</td>
<td>CamemBERT-base</td>
<td>None</td>
<td>28.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>further pre-trained</td>
<td>26.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MLM</td>
<td>23.38</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tri-training</td>
<td>17.11</td>
</tr>
<tr>
<td>fr</td>
<td>CamemBERT-base-ft-SQuAD2</td>
<td>None</td>
<td>47.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>further pre-trained</td>
<td>34.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MLM</td>
<td>46.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>tri-training</td>
<td>46.76</td>
</tr>
<tr>
<td>nl</td>
<td>RobBERT-base</td>
<td>None</td>
<td>34.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>further pre-trained</td>
<td>34.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MLM</td>
<td>42.84</td>
</tr>
<tr>
<td>nl</td>
<td>RobBERT-base-ft-SQuAD2</td>
<td>None</td>
<td>44.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>further pre-trained</td>
<td>34.61</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MLM</td>
<td>31.6</td>
</tr>
</tbody>
</table>

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.
<table>
<thead>
<tr>
<th>Language</th>
<th>Model</th>
<th>Setting</th>
<th>Dataset size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>fr</td>
<td>CamemBERT-base</td>
<td>None</td>
<td>28.75</td>
</tr>
<tr>
<td></td>
<td>CamemBERT-base-ft-SQuAD-fr</td>
<td>None</td>
<td><strong>47.3</strong></td>
</tr>
<tr>
<td></td>
<td>XLM-RoBERTa-base</td>
<td>None</td>
<td>23.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Simple</td>
<td>22.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MLM</td>
<td>33.36</td>
</tr>
<tr>
<td></td>
<td>XLM-RoBERTa-base-ft-SQuAD-fr</td>
<td>None</td>
<td>46.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Simple</td>
<td>46.08</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MLM</td>
<td>47.0</td>
</tr>
<tr>
<td>nl</td>
<td>RobBERT-base</td>
<td>None</td>
<td>34.62</td>
</tr>
<tr>
<td></td>
<td>RobBERT-base-ft-SQuAD-nl</td>
<td>None</td>
<td>44.05</td>
</tr>
<tr>
<td></td>
<td>XLM-RoBERT-base</td>
<td>None</td>
<td>33.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Simple</td>
<td>17.23</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MLM</td>
<td>37.66</td>
</tr>
<tr>
<td></td>
<td>XLM-RoBERT-base-ft-SQuAD-nl</td>
<td>None</td>
<td>43.73</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Simple</td>
<td>43.32</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MLM</td>
<td><strong>45.94</strong></td>
</tr>
</tbody>
</table>

Table 7: $F_1$ 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.
<table>
<thead>
<tr>
<th>Attribute</th>
<th>Question</th>
<th>#/Annotated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accused of crime</td>
<td>What crimes did the person commit?</td>
<td>107</td>
</tr>
<tr>
<td>Also known as</td>
<td>What other aliases does the person have?</td>
<td>103</td>
</tr>
<tr>
<td>Clothing</td>
<td>What clothes did the person wear?</td>
<td>656</td>
</tr>
<tr>
<td>Companions</td>
<td>What are the names of the person’s friends?</td>
<td>49</td>
</tr>
<tr>
<td>Contact address</td>
<td>Where does the contact person of the ad live?</td>
<td>740</td>
</tr>
<tr>
<td>Contact occupation</td>
<td>What does the contact of the ad do for a living?</td>
<td>278</td>
</tr>
<tr>
<td>Country marks</td>
<td>What country marks does the person have?</td>
<td>63</td>
</tr>
<tr>
<td>Destination (region)</td>
<td>What is the destination region of the person?</td>
<td>15</td>
</tr>
<tr>
<td>Destination (specified)</td>
<td>What is the name of the destination?</td>
<td>118</td>
</tr>
<tr>
<td>Disease</td>
<td>What kind of diseases does the person have?</td>
<td>91</td>
</tr>
<tr>
<td>Given name</td>
<td>What is the given name of the person?</td>
<td>693</td>
</tr>
<tr>
<td>Given surname</td>
<td>What is the last name of the person?</td>
<td>196</td>
</tr>
<tr>
<td>Injuries</td>
<td>How was the person injured?</td>
<td>63</td>
</tr>
<tr>
<td>Language</td>
<td>What are the communication skills of the person?</td>
<td>319</td>
</tr>
<tr>
<td>Literacy</td>
<td>What is the literacy level of the person?</td>
<td>8</td>
</tr>
<tr>
<td>Motivation</td>
<td>Why did the person escape his owner?</td>
<td>4</td>
</tr>
<tr>
<td>Name of contact</td>
<td>Who is the contact person for the ad?</td>
<td>678</td>
</tr>
<tr>
<td>Origin</td>
<td>Where does the person originate from?</td>
<td>28</td>
</tr>
<tr>
<td>Other reward</td>
<td>What other rewards were offered?</td>
<td>382</td>
</tr>
<tr>
<td>Owner</td>
<td>Who is the owner of the person?</td>
<td>395</td>
</tr>
<tr>
<td>Owner address</td>
<td>Where does the owner of the person live?</td>
<td>270</td>
</tr>
<tr>
<td>Owner occupation</td>
<td>What does the owner of the person do for a living?</td>
<td>78</td>
</tr>
<tr>
<td>Personality</td>
<td>What are the personality traits of the person?</td>
<td>15</td>
</tr>
<tr>
<td>Physical characteristics</td>
<td>What are the physical characteristics of the person?</td>
<td>568</td>
</tr>
<tr>
<td>Physical scars</td>
<td>What scars does the person have?</td>
<td>131</td>
</tr>
<tr>
<td>Plantation marks</td>
<td>What plantation marks does the person have?</td>
<td>23</td>
</tr>
<tr>
<td>Racial descriptor</td>
<td>What is the ethnicity of the person?</td>
<td>807</td>
</tr>
<tr>
<td>Ran from region</td>
<td>What is the name of the region the person escaped from?</td>
<td>3</td>
</tr>
<tr>
<td>Ran from specified</td>
<td>What is the name of the place the person escaped from?</td>
<td>406</td>
</tr>
<tr>
<td>Religion</td>
<td>What is the religion of the person?</td>
<td>13</td>
</tr>
<tr>
<td>Runaway date</td>
<td>What was the date of the event?</td>
<td>15</td>
</tr>
<tr>
<td>Skills</td>
<td>What is the set of skills of the person?</td>
<td>55</td>
</tr>
<tr>
<td>Specified occupation</td>
<td>What does the person do for a living?</td>
<td>98</td>
</tr>
<tr>
<td>Stutters</td>
<td>Does the person stutter?</td>
<td>22</td>
</tr>
<tr>
<td>Total reward</td>
<td>How much reward is offered?</td>
<td>780</td>
</tr>
</tbody>
</table>

Table 8: The attributes of the Runaways dataset
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.