Batavia asked for advice. Pretrained language models for Named Entity Recognition in historical texts.

Sophie Arnoult
Vrije Universiteit Amsterdam
s.i.arnoult@vu.nl

Lodewijk Petram
Huygens ING
lodewijk.petram@huygens.knaw.nl

Piek Vossen
Vrije Universiteit Amsterdam
p.t.j.m.vossen@vu.nl

Abstract

Pretrained language models like BERT have advanced the state of the art for many NLP tasks. For resource-rich languages, one has the choice between a number of language-specific models, while multilingual models are also worth considering. These models are well known for their crosslingual performance, but have also shown competitive in-language performance on some tasks. We consider monolingual and multilingual models from the perspective of historical texts, and in particular for texts enriched with editorial notes: how do language models deal with the historical and editorial content in these texts? We present a new Named Entity Recognition dataset for Dutch based on 17th and 18th century United East India Company (VOC) reports extended with modern editorial notes. Our experiments with multilingual and Dutch pretrained language models confirm the crosslingual abilities of multilingual models while showing that all language models can leverage mixed-variant data. In particular, language models successfully incorporate notes for the prediction of entities in historical texts. We also find that multilingual models outperform monolingual models on our data, but that this superiority is linked to the task at hand: multilingual models lose their advantage when confronted with more semantical tasks.

1 Introduction

Pretrained language models (Peters et al., 2018; Radford et al., 2018; Devlin et al., 2019) have recently advanced the state-of-the-art in many NLP tasks, providing deep contextual language representations and ease of deployment. BERT (Devlin et al., 2019) has proven particularly successful, combining the attention-based contextual representations of Transformers (Vaswani et al., 2017) with a simple fine-tuning procedure, allowing to deploy quickly to any sequence or document classification task.

BERT has given birth to a myriad of variants, differing by their training procedure, model size or language. Resource-rich languages in particular are spoilt for choice. Dutch for instance has two main monolingual models: the BERT-based BERTje (de Vries et al., 2019) and RobBERT (Delobelle et al., 2020), itself based on RoBERTa (Liu et al., 2019), a revision of BERT's training procedure. But multilingual models like mBERT (Devlin, 2018) and XLM-R (Conneau et al., 2020) are also applicable. These models, which are trained on 104, respectively 100 languages at once, perform well on crosslingual transfer (Pires et al., 2019; Wu and Dredze, 2019; Conneau et al., 2020). Pires et al. (2019) notably show that mBERT learns generic representations over different input languages and scripts. This makes multilingual models particularly interesting for historical texts, not only because these contain noncontemporary language, but also because they exhibit language variation due to language change over long periods of time or unstandardized spelling. But what when historical texts are accompanied by modern notes? How do monolingual and multilingual models compare on the historical and editorial parts of such texts, and on their combination?

If we only consider modern notes, it is not given that monolingual models outperform multilingual models on their training language. Notwithstanding possible differences in genre, the nature of the linguistic task at hand is an important factor in determining whether a monolingual or multilingual model is more appropriate. de Vries et al. (2019) and Delobelle et al. (2020) show for Dutch that mBERT is competitive with monolingual models on POS tagging (on Universal Dependencies POS), while Dutch models perform better at semantic tasks like Semantic Role Labelling or language-specific tasks like agreement resolution. For Named Entity Recognition (NER), results appear mixed, as both BERTje and RobBERT

Proceedings of LaTeCH-CLfL 2021, pages 21–30
Punta Cana, Dominican Republic (Online), November 11, 2021.

perform in-between mBERT and a version optimized by Wu and Dredze (2019). How semantic or language-specific does a task need to be for monolingual models to perform better than multilingual models in historical texts with editorial notes?

In this article, we consider these questions in the context of NER in 17th and 18th century Dutch texts. We introduce a new NER dataset based on the *General Letters of the United East India Company* (VOC)[1]. These official letters were written to the board of the VOC between 1610 and 1795. They report on the activities of the VOC (trade, conflicts, intelligence) in and around Indonesia and up to Japan and South Africa. Between 1960 and 2017, a selection of this corpus, spanning the time period 1610-1767, was transcribed and edited by the Huygens Institute for the History of the Netherlands (Huygens ING) and its predecessors. Annotating entities in these letters is part of a larger effort to facilitate historical research on these texts.

Our dataset[2] consists of 24.5k entities, with a repartition of 40%/60% between historical text and editorial notes, where the historical text covers the period 1621 to 1760. We introduce a semantic component in this dataset by distinguishing metonymical uses of locations (which represent almost half of all entities). This allows us to compare two NER variants: a standard one, where locations form a single entity type; and a more semantic one, where models additionally must distinguish locations in agent-like roles, as in for instance, *Batavia asked for advice*.

We compare the performance of monolingual and multilingual models on both NER variants through BERTje (de Vries et al., 2019), RobBERT (Delobelle et al., 2020), mBERT (Devlin, 2018) and XLM-R (Conneau et al., 2020). On standard NER, we find that multilingual models outperform monolingual models on both historical text and modern notes. In crosstext finetuning and evaluation, we confirm the strong crosslingual transfer ability of multilingual models in the context of historical language variation. We also show that notes and historical text are complementary for NER, as all pretrained models, multilingual *and* monolingual, benefit from their combination. On the more semantic-oriented NER task, we find that monolingual and multilingual models perform competitively, with RobBERT and mBERT performing best. This confirms the importance of the semantic nature of a task as a factor in choosing between monolingual and multilingual models.

[1] http://resources.huygens.knaw.nl/vocgeneralemissiven/index_html_en

[2] Dataset and code are available at https://github.com/cltl/voc-missives

2 Annotations

The labelset consists of five base labels—LOC, ORG, PER, REL and SHP—for the entity types *locations*, *organisations*, *persons*, *religions* and *ships*. Following (Benikova et al., 2014), this labelset is extended with secondary labels of the form X_{deriv} or X_{part} for expressions derived from entity names by grammatical derivation (as with for instance, the location *Banda* and the derived *Bandanezen*[3]) or composition (as with *Java* and *Javakoffie*[4]). The labelset is extended with a GPE label (*geopolitical entities*) to distinguish metonymical use of location names.

2.1 Data selection

The data selected for annotations consist of letters spread out through the *General Letters* corpus[5]. In editing these letters, Huygens ING transcribed parts of the letters, and summarized other parts, which appear as in-text paragraph notes in the digital version of the corpus[6]. Editorial comments appear as footnotes.

We use the post-OCR version of the corpus[7] made available in Text-Fabric (Roorda, 2016, 2019). This is a clean text overall, with some errors in character recognition (e.g., *Ib* instead of *lb*, *S* Malo* instead of *St Malo*), page-to-text formatting (*Chris -toffel* for *Christoffel*) or tokenization (*Schippers ,* instead of *Schippers,*).

We selected 25 letters for annotation, keeping the historical text and editorial notes of these letters apart to facilitate their separate use. Our final dataset consists of 43 documents: 22 historical *text* documents and 21 editorial *notes* documents[8].

[3] Bandanese

[4] Java coffee

[5] The original corpus contains fourteen volumes, the first thirteen of which were available for annotations.

[6] http://resources.huygens.knaw.nl/retroboeken/generalemissiven/#page=0&accessor=toc&view=homePane

[7] https://github.com/annotation/app-missieven, version 0.8.1

[8] A number of documents were unfortunately lost because of server errors. From the 25 letters, 18 are annoted for both *text* and *notes*, 4 for *text* only, and 3 for *notes* only.

	text	notes	total
tokens	233k	201k	434k
entities	9.57k	14.9k	24.5k
density (%)	4.1	7.4	5.6
entity types			
LOC	3927	6525	10.5k
LOC_{deriv}	2227	1414	3641
LOC_{part}	14	21	35
GPE	43	854	897
ORG	994	995	1989
ORG_{part}	8	41	49
PER	1665	3337	5002
REL	10	4	14
REL_{deriv}	159	73	232
REL_{part}	7	10	17
SHP	520	1652	2172

Table 1: Annotation counts

2.2 Annotation counts

Summary statistics are provided in Table 1. Compared to the historical text, the editorial notes show a higher density of entities per token, and they refer less to peoples (through derived forms of *locations* and *religions*), and more to *locations*, *persons* and *ships*. This agrees with notes commenting on *primary* named entities. In contrast, the skewed distribution of *geopolitical entities* between text and notes appears to be a stylistic artefact: metonymical use of locations is in fact concentrated in three notes documents (from volumes 11 and 12), which contain 88.0% of all *geopolitical entities* (for 14.8% of all entities and 11.3% of all tokens). While we consider here the editorial notes as being linguistically homogeneous, their writing spanned more than fifty years and they are at least stylistically varied.

2.3 Guidelines

Our annotations focus on named entities and their derived forms, aiming at annotations that are both consistent for NER and adaptable for extensions. We present here the main considerations in annotating different types of entities.

Locations Only the named entity is marked as an entity in compositional location names (*rif van* $[Luang]_{LOC}$), except when the full expression is treated as a given name. This is in particular the case with coastal areas, which were often the only known part of an island for the VOC:

(1) [Java's Oostkust]$_{LOC}$, [Sumatra's Westkust]$_{LOC}$, [Malabar]$_{LOC}$ en [Ceylon]$_{LOC}$ [9]

The *Coast* then refers to a specific coastal area, like in this note:

(2) de [Custe]$_{LOC}$ is hier en elders de [Kust van Coromandel]$_{LOC}$[10]

Practically, we annotate compositional location names as entities when their constituents are capitalized:

(3) expeditie ter [Oostkust van Java]$_{LOC}$[11]

(4) nabij de noordkust van [Java]$_{LOC}$[12]

Derived (adjectival) forms of locations are annotated with LOC_{deriv} and composed names with LOC_{part}:

(5) de [Bandanezen]$_{LOC_{deriv}}$

(6) Sijn [Portugeesche]$_{LOC_{deriv}}$ Majesteyt

(7) [Javakoffie]$_{LOC_{part}}$

Location names with a semantic role of *agent*, *theme*, *experiencer*, *benefactive*, or of *trading* or *political actor* are marked with a distinct GPE label (we use this label for our semantic-oriented NER experiments; GPE entities are relabelled as LOC for standard NER experiments).

(8) [Batavia]$_{GPE}$ heeft advies gevraagd[13]

(9) er is aan [Coromandel]$_{GPE}$ opgedragen[14]

(10) de inkomsten van [Malakka]$_{GPE}$[15]

(11) de vrede met [Frankrijk]$_{GPE}$[16]

[9] *Java's Eastcoast, Sumatra's Westcoast, [...]*
[10] *The Coast refers here and elsewhere to the Coast of Coromandel*
[11] *expedition to the Eastcoast of Java*
[12] *near the north coast of Java*
[13] *Batavia has asked for advice*
[14] *Coromandel has been ordered to*
[15] *the revenues of Malakka*
[16] *peace with France*

Organisations Compositional organisation names are annotated as a whole in principle (*Rade van India*, *College van Schepenen*). An exception is made for *Kamer* and *Compagnie* which are used productively in the context of the VOC:

(12) de [kamers]$_{ORG}$ [Delft]$_{LOC}$, [Rotterdam]$_{LOC}$ en [Hoorn]$_{LOC}$

(13) de respective [Comp.en]$_{ORG}$ van [Engelant]$_{LOC}$ ende [Nederlant]$_{LOC}$

(14) wegen d'[Engelse]$_{LOCderiv}$ en de [Nederlantse]$_{LOCderiv}$ [Comp.en]$_{ORG}$

(15) de [France]$_{LOCderiv}$ [Comp.ie]$_{ORG}$ van [S* Malo]$_{LOC}$

Common nouns denoting organisations are not annotated[17]:

(16) het [Siamse]$_{LOCderiv}$ hoff[18]

Persons Person names are annotated without qualifiers or titles[19]:

(17) ingenieur [Albert Legrand]$_{PER}$

(18) Radja [Simorang]$_{PER}$

Religions Religion names and derived forms are a small but relevant group of entities in the context of the VOC:

(19) concurrentie van [Engelsen]$_{LOCderiv}$ en [Moren]$_{RELderiv}$[20]

We systematize the annotation of religious groups by including non-capitalized names:

(20) De [heidense]$_{RELderiv}$ bewoners [...] de [Moslimse]$_{RELderiv}$ kustbewoners[21]

Ships Ships form a considerable part of the annotations. They are annotated without determiners or qualifiers:

(21) de [Loenderveen]$_{SHP}$

[17] These are historically relevant while falling out of a linguistic definition of named entities. We might extend *organisations* to terms like *hof*, *gouvernement* or *comptoir* in a future version of the dataset.

[18] *the Siamese court*

[19] Following (Benikova et al., 2014). The time and space covered by the *General Letters* corpus entails however a great variety of personal titles, that can be hard to tell from person names. We might extend our labelset with titles in the future.

[20] *competition from Englishmen and Moors*

[21] *the pagan inhabitants [...] the Moslim coastal inhabitants*

2.4 Annotation process

Annotations were performed with Inception (Klie et al., 2018) by two annotators, following guidelines inspired from (Benikova et al., 2014). To disambiguate difficult cases, the annotators could rely on the indices of persons, locations and ships accompanying each volume of the corpus, as well as on a glossarium[22]. Annotations were performed for the most part on raw text[23] with the help of a gazetteer compiled from the indices; a few documents were preannotated with either string matching from the gazetteer or with a preliminary NER system.

Agreement was measured halfway through the annotation process, on three documents (two with historical text and one with editorial notes). We measure inter-annotator agreement with F score: like Brandsen et al. (2020) and Deleger et al. (2012), we question the use of Cohen's kappa for NER, as it is unclear how chance agreement should be defined for NER.

Table 2 provides F scores for the text and notes and details cases of agreement and disagreement between both annotators. To this end, we first pair up annotations with a same span to isolate cases of agreement and cases of *label* disagreement. We then attempt to pair remaining annotations, and count annotations that overlap with annotations of the other annotator; those that do are cases of *span* disagreement, while the remaining cases are mentions that only one of the annotators identifies as *entities*.

The analysis of disagreements presented next revealed some inherent difficulties with annotating historical texts, errors in the annotations, but also unclarities in the guidelines. We used this analysis to streamline guidelines and correct annotations[24].

Disagreement analysis Most cases of label disagreement result from the confusion between person names and location or ship names, or between location names and derived forms thereof. Person names cannot always be distinguished from location or ship names by the linguistic context alone, especially with earlier language variants. In this example for instance:

[22] http://resources.huygens.knaw.nl/vocglossarium/VocGlossarium/zoekvoc

[23] Extracted from TEI post-OCR files. Annotations on this text were later ported to the Text-Fabric release of the corpus.

[24] Corrections were performed by one of the authors, after annotations were gathered.

	text	notes	total
F1	88.5	92.1	90.8
entities	492/484	910/877	1402/1361
agreeing	432	823	1255
disagreeing			
- label	10	24	34
- span	39/39	24/25	63/64
- entity	11/3	39/5	50/8

Table 2: Inter-annotator agreement. Pairs of counts correspond to the respective annotation counts of each annotator.

(22) vroeg de sengadji van Lamakera Mauwadasje[25]

the indices tell us that *Lamakera* refers to a location and *Mauwadasje* to a person (and the glossarium that *sengadji* denotes the head of a village or district). Linguistic context alone is not enough to distinguish person from location, and the form of the apposition of *Mauwadasje* is also unusual for contemporary Dutch.

Most cases of span disagreement[26] concern infix abbreviations like *Comp.e* for *Compagnie* and qualifiers like *Edele* (*noble*) for *Compagnie*, person titles like *Khan* or *Radja* and location qualifiers as in *engte van Pambenaar*[27] or *Noord-Celebes*.

Most cases of entity disagreement are due to omissions. Other cases concern whether or not to annotate: political actors in general like *gouvernement*; derived forms of location names not denoting peoples, as in *het Engels schip*; compositional location names like *Oostkust*; metonymical uses of locations[28].

3 Experimental Setup

3.1 Models

BERTje (de Vries et al., 2019) is a Dutch model trained on 12GB of data from mixed, mainly contemporary sources. The model is structurally equivalent to BERT$_{base}$, with 12 Transformer layers of size $H = 768$ and 12 attention heads. The tokenizer is based on Sentence-Piece (Kudo and Richardson, 2018), and has a vocabulary size of 30k. Unlike BERT, BERTje is trained on a Sentence-Order Prediction objective next to Masked Language Modelling (MLM).

RobBERT (Delobelle et al., 2020) is a Dutch model trained on 39GB of data from the OSCAR corpus (Ortiz Suárez et al., 2019). The model is structurally equivalent to BERT$_{base}$ while following the training procedure of RoBERTa (Liu et al., 2019), with a dynamic MLM objective, and a tokenizer based on byte-level BPE following (Radford et al., 2019), with a vocabulary size of 40k (we consider RobBERT v2).

mBERT (Devlin, 2018) is a multilingual model trained on Wikipedia dumps of the 104 languages best represented on Wikipedia (we consider the *cased* version); the model is trained without knowledge of which languages are input, while data are sampled to compensate for less represented languages. The vocabulary is shared across languages and built with Word-Piece (Wu et al., 2016) and has a size of 110k. The model is structurally equivalent to BERT$_{base}$, and is trained with an MLM and Next Sentence Prediction objective.

XLM-R (Conneau et al., 2020) is a multilingual model trained on a CommonCrawl corpus of 2.5TB, with a vocabulary of 250k tokens built with the Sentence-Piece tokenizer. Like RoBERTa (and RobBERT), XLM-R is trained with an MLM objective, taking sequences of tokens of fixed length as input instead of sentences. We use XLM-R$_{base}$, which is structurally equivalent to BERT$_{base}$.

3.2 Data

The data are split in train/dev/test sets so as to obtain comparable splits for the notes and historical text while keeping the notes and text of a same letter together. Besides, we selected the test set for the historical text from the earliest part of the corpus, so as to make it more challenging for a model trained and validated on later text; as the earlier letters contain little notes, the test set for the notes is complemented with a letter for which only the notes were annotated. Data selection for

[25] *asked the sengadji of Lamakera Mauwadasje*

[26] Annotation counts differ for the notes because one entity (*Castor en Pollux*) was marked as two entities by the second annotator.

[27] *Pambenaar's strait*

[28] These were first marked by double LOC/ORG labels. In these cases, the annotators would disagree on adding an ORG label.

training	1658-1730, 1743-44, 1759-60
validation	1731-40, 1741^t, 1750^n
test	1621-43, 1647^n, 1752^n

Table 3: Year ranges of selected letters for each data subset. t: text only; n: notes only

	text	notes	all
training			
tokens	170k	160k	330k
entities	7192	11.5k	18.7k
avg. length	56.6	89.4	68.9
validation			
tokens	30.3k	24.4k	54.7k
entities	981	1991	2972
avg. length	106.8	71.9	87.8
test			
tokens	32.3k	16.7k	49.0k
entities	1401	1436	2837
avg. length	142.9	32.5	66.3

Table 4: Data split. Average sequence lengths are in tokens, before subword tokenization.

the notes also accommodates for GPE labels being concentrated in three letters (from 1744, 1750 and 1752). Data selection is summarized in Table 3.

The data are tokenized with the IXA-pipe tokenizer (Agerri et al., 2014). We do not segment sentences with the tokenizer, but take paragraphs and separate notes as text units, splitting sequences longer than 256 subword units for each language model. This is motivated practically by the tokenizer being to greedy (the letters abound with abbreviations), but we also believe that working at the paragraph level may be beneficial as it provides more context for NER. Summary statistics are provided in Table 4.

For both standard and semantic-oriented experiments, rare labels are mapped to lexically-related labels: REL and REL$_{part}$ to REL$_{deriv}$, LOC$_{part}$ to LOC$_{deriv}$, ORG$_{part}$ to ORG. Additionally, GPE labels are remapped to LOC for standard NER.

3.3 Hyperparameter settings

All models are fine-tuned with the HuggingFace Transformers library (Wolf et al., 2020). Adam parameters are: $\beta_1 = 0.9$; $\beta_2 = 0.999$, $\epsilon = 10^{-8}$. We used training batch sizes of 16, a learning rate of $5 \cdot 10^{-5}$ and no weight decay. We did not perform hyperparameter search, but we fine-tuned all models with three different seeds (1, 10 and 100), and we report average values with standard deviation across the three runs. For each run, we selected the best model within 4 epochs, taking loss on the validation set as a criterion and keeping checkpoint models every 100 steps when fine-tuning on the text or notes only, and every 200 steps when fine-tuning on all the data[29]. All experiments were run on a single Tesla P100 GPU. Fine-tuning on all the data takes around 15 minutes per run for all models.

4 Results

We first present results on standard NER, comparing the in-text (text or notes) and crosstext performance of monolingual and multilingual models, and assessing the reciprocal contribution of text and notes. We end with results on semantic-oriented NER.

4.1 In-text performance

Table 5 presents results with fine-tuning and evaluation on the same part of the dataset (text, notes or both). Multilingual models perform better than monolingual models on average, albeit with a smaller margin for the notes and the entire dataset. Results are higher for all models on the notes than on the text; this may have to do with a more homogeneous language in the notes, but also with differences in sequence lengths, which are long on average in the text's testset and short in the notes[30]. The length of text sequences coupled with our working at the paragraph level may also explain the higher unstability of results of BERTje and mBERT, as these were trained on the sentence level.

4.2 Crosstext performance

Crosstext results are presented in Table 6. These confirm the superiority of multilingual models in crosslingual transfer: while all models fine-tuned on text loose a few points when predicting on notes,

[29] Fine-tuning takes about 550 steps for the notes, 850 for the text and 1400 for all the data.

[30] This resulted by chance from our data split. While it might be interesting to try another split resulting in more even sequence lengths, the other constraints we followed, combined with the limited number of letters, limit possibilities in that direction.

	text	notes	all
BERTje	84.4 (2.5)	91.8 (0.4)	90.4 (0.5)
RobBERT	86.1 (0.8)	92.7 (0.5)	91.1 (0.9)
mBERT	**88.7** (3.2)	**93.4** (0.7)	92.2 (0.6)
XLM-R$_{base}$	88.2 (0.6)	**93.4** (0.4)	**92.3** (0.3)

Table 5: In-text fine-tuning and evaluation (F1 scores and standard deviation)

		text	notes
BERTje	text	84.4 (2.5)	82.8 (2.9)
	notes	68.7 (3.3)	91.8 (0.4)
RobBERT	text	86.1 (0.8)	84.1 (1.5)
	notes	71.9 (1.8)	92.7 (0.5)
mBERT	text	88.7 (3.2)	**86.8** (3.6)
	notes	**79.5** (1.2)	93.4 (0.7)
XLM-R$_{base}$	text	88.2 (0.6)	84.7 (0.6)
	notes	77.9 (2.9)	93.4 (0.4)

Table 6: Crosstext fine-tuning and evaluation. Models fine-tuned on the text or notes (rows) and evaluated on text and notes (columns).

		in-text	all
BERTje	text	84.4 (2.5)	87.8 (0.8)
	notes	91.8 (0.4)	92.9 (0.6)
RobBERT	text	86.1 (0.8)	89.3 (1.1)
	notes	92.7 (0.5)	92.9 (0.7)
mBERT	text	88.7 (3.2)	90.8 (0.7)
	notes	93.4 (0.7)	93.6 (0.6)
XLM-R$_{base}$	text	88.2 (0.6)	91.0 (0.5)
	notes	93.4 (0.4)	93.6 (0.1)

Table 7: Reciprocal contribution of notes and text: fine-tuning on in-text data (text or notes) or on all data, and predicting on text or notes (rows).

	text		notes	
	in-text	all	in-text	all
LOC	93.0 (0.6)	94.4 (0.4)	96.7 (0.3)	96.7 (0.1)
LOCd	92.0 (0.3)	93.1 (0.1)	89.6 (0.4)	92.5 (1.0)
ORG	87.5 (1.9)	90.2 (1.6)	89.8 (1.3)	90.9 (2.2)
PER	74.1 (4.2)	79.3 (2.2)	89.4 (1.2)	88.4 (1.1)
RELd	92.3 (1.9)	92.1 (2.3)	39.1 (36.7)	90.0 (5.8)
SHP	60.0 (5.0)	76.2 (1.0)	86.8 (0.4)	86.4 (1.5)
all	88.2 (0.6)	91.0 (0.5)	93.4 (0.4)	93.6 (0.1)

Table 8: Detailed contribution of notes and text fine-tuning data to text and notes predictions, for XLM-R$_{base}$.

multilingual models fine-tuned on notes are more robust when predicting on text, their performance dropping then by about 15 points compared to 20 points or more for monolingual models.

4.3 Reciprocal contribution of notes and text

Table 7 shows the effect of adding out-of-text data for fine-tuning. We see that notes are informative for NER in historical text, as performance on the text increases by 2 to 3 points when fine-tuning on all data. The reverse does not hold: performance on the notes also increases when fine-tuning on all data, but only minimally (by 0.2 point for all models except BERTje, for which scores improve by 1.1 point). These trends hold for all models, monolingual or multilingual, pointing to the general contextual value of notes for named entities combined with a general ability of pretrained language models to exploit context—while monolingual models are trained on a single language, they retain the ability to adapt to a variety of language variants in fine-tuning.

Table 8 details the contribution of notes and text per label for XLM-R. Notes are informative for all entities except *religions* and derived forms, which are poorly represented in the notes (results for the REL$_{deriv}$ label are also very unstable in the notes, as witnessed by the high standard deviation). The highest gains are obtained for *persons* and *ships*, which are also much better represented in the notes than the text. But notes are also informative for entities that are well represented in the text (*locations*, derived forms thereof and *organisations*). Reciprocally, the text is informative for entities that are comparatively better represented: derived forms of *locations*, *organisations* and *religions*.

4.4 Locations as political actors

We end with a semantic oriented NER experiment, where models additionally must distinguish between metonymical (GPE) and standard uses (LOC) of locations. Results are presented in Table 9. Adding this semantic orientation makes it

	-GPE	+GPE
BERTje	90.4 (0.5)	88.4 (0.5)
RobBERT	91.1 (0.9)	**89.0** (0.5)
mBERT	**92.2** (0.6)	**89.1** (1.2)
XLM-R$_{base}$	**92.3** (0.3)	88.4 (2.2)

Table 9: Distinguishing metonymical uses of locations. -GPE: no distinction; +GPE: distinction between LOC and GPE labels. Models fine-tuned and evaluated on all the data. Boldface marks the two best performing models per case.

	mBERT		RobBERT	
	-GPE	+GPE	-GPE	+GPE
GPE	-	82.0 (0.3)	-	84.0 (1.5)
LOC	95.7 (0.6)	93.0 (0.8)	94.8 (0.5)	92.2 (0.7)
LOCd	92.7 (0.4)	91.0 (1.3)	92.2 (0.4)	91.1 (1.1)
ORG	90.6 (3.4)	90.5 (1.5)	88.6 (1.7)	88.7 (1.8)
PER	84.7 (0.6)	84.0 (1.6)	83.8 (0.4)	82.7 (1.4)
RELd	93.4 (3.9)	78.1 (23.8)	91.3 (4.1)	92.8 (1.9)
SHP	82.1 (1.9)	77.0 (1.6)	79.6 (4.4)	80.6 (0.3)
all	92.2 (0.6)	89.1 (1.2)	91.1 (0.9)	89.0 (0.5)

Table 10: Distinguishing metonymical uses of locations: detailed label scores.

a harder task for all models. Monolingual models however perform relatively better, with RobBERT almost equalling mBERT as best performing model. Multilingual models suffer a larger drop in performance than monolingual models, coupled with more instable results.

We compare detailed label scores for RobBERT and mBERT in Table 10. GPE-label distinction negatively affects prediction of all entity types for mBERT. In contrast, RobBERT performs better than mBERT on GPE prediction, and benefits from the GPE/LOC distinction for the prediction of a few entity types (derived forms of *religions* and *ships*). We observe similar trends with XLM-R$_{base}$ and BERTje, respectively (for BERTje, scores improve for REL$_{deriv}$ and LOC$_{deriv}$ by 1 point).

5 Conclusion

We have introduced a new Dutch dataset for Named Entity Recognition, consisting of Early Modern Dutch VOC letters and modern editorial notes. Comparing monolingual and multilingual pretrained language models, we confirm the stronger crosslingual abilities of multilingual models, while showing that both monolingual and multilingual models can leverage on mixed language variants at fine-tuning. For Named Entity Recognition, pretrained language models are notably able to leverage on notes to improve text predictions. We have further shown that multilingual pretrained language models not only perform better on the historical part of this dataset, but also on modern Dutch notes. However, the superiority of multilingual models on modern Dutch is very much linked to the task at hand, as orienting the data to more semantic distinctions can turn the tables for monolingual models.

Acknowledgements

This work is part of a *Clariah Work Package 6: Text* project, and supported by NWO project number CP-WP6-19-005. We thank the members of the *VOC use case* group for interesting discussions and the anonymous reviewers for their pertinent comments.

References

Rodrigo Agerri, Josu Bermudez, and German Rigau. 2014. IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2524–2531, Reykjavik, Iceland. European Language Resources Association (ELRA).

Alex Brandsen, Suzan Verberne, Milco Wansleeben, and Karsten Lambers. 2020. Creating a Dataset for Named Entity Recognition in the Archaeology Domain. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4573–4577, Marseille, France. European Language Resources Association.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. 2019. BERTje: A Dutch BERT Model. ArXiv: 1912.09582.

Louise Deleger, Qi Li, Todd Lingren, Megan Kaiser, Katalin Molnar, Laura Stoutenborough, Michal Kouril, Keith Marsolo, and Imre Solti. 2012. Building Gold Standard Corpora for Medical Natural Language Processing Tasks. *AMIA Annual Symposium Proceedings*, 2012:144–153.

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. 2018. Multilingual BERT release. Original-date: 2018-10-25T22:57:34Z.

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.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. 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: 1907.11692.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*, Cardiff, United Kingdom. Leibniz-Institut für Deutsche Sprache.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. Technical report, OpenAI.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Dirk Roorda. 2016. Text-Fabric. Software. Version 8.3.4.

Dirk Roorda. 2019. Text-fabric: handling Biblical data with IKEA logistics. *HIPHIL Novum*, 5(2):126–135.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. ArXiv: 1706.03762.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Łukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith

Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. ArXiv: 1609.08144.