Measuring Spurious Correlation in Classification:
"Clever Hans" in Translationese

Angana Borah[1], Daria Pylypenko[1], Cristina España-Bonet[2], Josef van Genabith[1,2]

[1]Saarland Informatics Campus, Saarland University, Germany
[2]DFKI GmbH, Germany

`anganaborah9@gmail.com, daria.pylypenko@uni-saarland.de,`
`{cristinae,josef.van_genabith}@dfki.de`

Abstract

Recent work has shown evidence of "Clever Hans" behavior in high-performance neural translationese classifiers, where BERT-based classifiers capitalize on spurious correlations, in particular topic information, between data and target classification labels, rather than genuine translationese signals. Translationese signals are subtle (especially for professional translation) and compete with many other signals in the data such as genre, style, author, and, in particular, topic. This raises the general question of how much of the performance of a classifier is really due to spurious correlations in the data versus the signals actually targeted for by the classifier, especially for subtle target signals and in challenging (low resource) data settings. We focus on topic-based spurious correlation and approach the question from two directions: (i) where we have no knowledge about spurious topic information and its distribution in the data, (ii) where we have some indication about the nature of spurious topic correlations. For (i) we develop a measure from first principles capturing alignment of unsupervised topics with target classification labels as an indication of spurious topic information in the data. We show that our measure is the same as purity in clustering and propose a "topic floor" (as in a "noise floor") for classification. For (ii) we investigate masking of known spurious topic carriers in classification. Both (i) and (ii) contribute to quantifying and (ii) to mitigating spurious correlations.

1 Introduction

The term *translationese* refers to systematic linguistic differences between originally authored texts and translated texts in the same language (Gellerstam, 1986). Important aspects of translationese have been identified in the linguistic literature (Toury, 1980; Baker et al., 1993; Teich, 2012; Volansky et al., 2013), including source language interference, over-adherence to target language, simplification, explicitation, and implicitation. Translationese may manifest itself at lexical, syntactic, semantic, and discourse-related levels of linguistic description. While translationese signals are subtle (especially for professional human translation), corpus-based linguistic methods (Baker et al., 1993) and machine learning based classification methods (Volansky et al., 2013; Rabinovich and Wintner, 2015; Rubino et al., 2016; Pylypenko et al., 2021) are able to reliably distinguish between original and translated texts in the same language, genre, and style. While basic research focuses on identifying and categorizing aspects of translationese, research has also shown that translationese clearly impacts practical cross-lingual tasks that involve translated data (Singh et al., 2019; Zhang and Toral, 2019; Clark et al., 2020; Artetxe et al., 2020). Finally, translationese is sometimes regarded as (one of) the final frontier(s) of high-resource machine translation (Freitag et al., 2020, 2019; Ni et al., 2022).

In this paper, we focus on translationese classification (into original O and translated T data) using machine learning based approaches. Early work on translationese classification focused on manually engineered and linguistically inspired sets of features (n-grams, POS, discrete LM-based features etc.), using supervised classification models such as decision-trees or support vector machines (SVMs) (Baroni and Bernardini, 2005; Volansky et al., 2013; Rubino et al., 2016).

More recently, research focused on feature-and-representation learning neural network methods for translationese classification (Sominsky and Wintner, 2019; Pylypenko et al., 2021). Pylypenko et al. (2021) show that BERT-based approaches outperform handcrafted feature and SVM-based approaches by a large margin (15-20 accuracy points absolute). Amponsah-Kaakyire et al. (2022) show

that this performance difference is due to learned features (rather than the classifiers).

Using Integrated Gradient (IG) based input attribution methods (Amponsah-Kaakyire et al., 2022) also show that BERT (Devlin et al., 2019a) sometimes exploits topic differences between O and T data as spurious correlations with the target classification labels (original O and translation T) as short cuts, rather than "true" translationese signals: the T part of the (Europarl-based) data, translations from Spanish into German, happens to contain mentions of Spanish locations while German originals O tend to mention German location names. Spurious correlations in the data with target classification labels may cause "Clever Hans" behavior (Lapuschkin et al., 2019; Hernández-Orallo, 2019), where the classifier picks up accidental patterns in the data correlated with but otherwise unrelated to the classification target, in the case at hand, topic/content differences rather than proper linguistic indicators of translationese.

To the best of our knowledge, to date, we do not know how to measure spurious correlations between topic signals in the data and target classification labels, such as translationese (O and T). At the same time, this is an important question, as an answer would allow us to better understand to what extent we can trust a classifier to pick up on information truly relevant to the target classification labels, and to which extent a classifier is exploiting "Clever Hans", i.e. spurious correlations in the data with the target labels. This is especially pertinent with subtle classification targets and challenging low-resource data settings, as in translationese classification: translationese data sets tend to be small, and translationese signals are subtle while competing with many other signals in the data.

We approach our research question of "measuring spurious topic correlation in the data with respect to target classification labels" from two opposing ends: (i) where we assume no prior knowledge about topics in the data and (ii) where we have some idea about spurious topic signals in the data. We refer to (i) as **Chasing Unknown Unknowns** (Section 4 below)[1] and (ii) as **Chasing Known Unknowns** (Section 5 below). For (i) we use unsupervised topic modeling (LDA and BERTopic) and

we develop a measure from first principles that captures the alignment of (unsupervised) topics with target classification labels. Based on this we propose the concept of a "topic floor" in classification, akin to the concept of a "noise floor" in Electronic Engineering. We show that our alignment-based measure is the same as purity with respect to target classes in clustering. Given data, target classification labels and unsupervised topic models, our measure and noise floor provide an upper bound on how much spurious topic information may account for target classification labels. For (ii) we use masking of already identified spurious topic information, such as location names, in the data and measure classification accuracy with masked and unmasked versions of the data, to quantify the impact of the identified source of spurious correlation.

Our main contributions include the following:

1. We present a measure that, given a data set and target classification labels, quantifies the possible impact of *unknown spurious topic information* on classification. The measure is based on aligning unsupervised topics with the target labels. Based on this we propose the concept of a "topic floor" (akin to "noise floor" in Electronic Engineering) in classification.

2. We use masking to both quantify and mitigate *known spurious topic information.*

3. We present empirical results for topic floor and masking to quantify "Clever Hans" in the translationese data of Amponsah-Kaakyire et al. (2022). We use IG attribution to show that in masked settings where known spurious correlations are mitigated, BERT learns features closer to proper translationese.

2 Related Work

Puurtinen (2003); Ilisei et al. (2010); Volansky et al. (2013); Rabinovich and Wintner (2015); Rubino et al. (2016); Pylypenko et al. (2021) train classifiers to distinguish between originally authored and translated data. Many of them explore handcrafted and linguistically inspired feature sets, manual feature engineering, and a variety of classifiers including Decision Trees and Support Vector Machines (SVMs) and use feature ranking or attribution methods to reason back to particular dimensions of translationese and their importance in the data and classification results. Feature engineering

---

[1]Readers may relate this to a 2002 hearing with the then US Secretary of Defence Donald Rumsfeld. In one scenario we do not know the topics and their impact, the unknown unknowns; in the other we have some indication about a spurious topic but again do not know its impact, the known unknowns.

based translationese classification used SVM feature weights (Avner et al., 2016; Pylypenko et al., 2021), decision trees or random forests (Ilisei et al., 2010; Rubino et al., 2016), training separate classifiers for each individual feature (or feature sets) and comparing accuracies (Volansky et al., 2013; Avner et al., 2016), to explain results.

More recent research uses feature and representation learning approaches (sometimes augmented with hand-crafted features) based on neural networks (Sominsky and Wintner, 2019; Pylypenko et al., 2021). Pylypenko et al. (2021) shows that feature learning based approaches (e.g. a pretrained BERT based classifier) outperform hand-crafted and feature engineering based approaches (SVM) by as much as 15 to 20 percentage points absolute in classification accuracy. Amponsah-Kaakyire et al. (2022) show that the difference in classification accuracy is due to feature learning rather than the classifiers, and, using Integrated Gradients (IGs) (Sundararajan et al., 2017), provide evidence that the feature learning methods exploit some spurious correlations with the classification labels in the data, that are clearly not translationese, but topic related cues: the data are German originals O and translations T (into German) from Spanish, and Spanish place names are highly IG-ranked, given the trained classifiers.

Dutta Chowdhury et al. (2022) use divergence from isomorphism based graph distance measures to show that translationese is visible even in O and T word embedding spaces. While POS features have been used in feature-engineering-based translationese classification, some experiments in (Dutta Chowdhury et al., 2022) use POS (instead of the surface words) to mitigate possible topic influences on the graph divergence results. This is an approach we develop further (e.g. in terms of partial masking using NEs) in our work below.

3 Data

Our experiments use the monolingual German dataset from the Multilingual Parallel Direct Europarl (MPDE) corpus (Amponsah-Kaakyire et al., 2021) consisting of 42k paragraphs, with half of the paragraphs German (DE) originals (below we call this O) and the other half translations (below we call this T) from Spanish (ES) to German. The average length (in terms of tokens) per training example is 80. Like all of MPDE, the DE-ES subset contains only data from before 2004, since for

post-2004 data, it may not be known whether or not the source language SL is already the result of a translation (Bogaert, 2011). While this limits the amount of data, it ensures that the O and T data are clearly identifiable and "pure" O or T. Both O and T are German, but T is German translated from Spanish, and both coming from MPDE ensure that they are the same Europarl genre and style.

4 Chasing Unknown Unknowns

In this section, we assume that we have no prior information about topics and their distribution in the data. Because of this, we use unsupervised topic modeling. We develop a measure that checks whether, and if so to what extent, the topics established in this fashion *align* with the target classes O and T in our data. The measure quantifies to which extent topic is a giveaway for translationese.

4.1 How to Measure Topic Bias Relevant to Translationese Classification?

The goal is to investigate the amount and distribution of topic signal in O and T data, that could be used as a spurious signal in translationese (i.e. O and T) classification. As initially we do not know anything about possible topics and their distribution in the data, we use standard approaches to unsupervised topic modeling, like Latent Dirichlet Allocation (LDA) (Blei et al., 2001) and BERTopic (Grootendorst, 2022). Both LDA and BERTopic will cluster our data into classes, i.e. topics. How can we measure whether the topics established by the topic model are potentially relevant to translationese classification? Topics are relevant to O and T translationese classification if the paragraphs in each of the topics are either mostly O paragraphs or if they are mostly T paragraphs, in other words if topics are well *aligned* to either O or T. If this is the case, a translationese classifier may learn to use topic, rather than proper translationese signals (or a mix of both) in translationese classification. To give a simple (and extreme) example, suppose we take the union[2] of O and T and cluster the union using LDA into, say, two topics (classes) top_1 and top_2. If (and this is the extreme case) $top_1 = O$ and $top_2 = T$ (or vice versa, i.e. $top_1 = T$ and $top_2 = O$), then topic perfectly predicts O and T. We would like our measure to capture this, and we would like the measure to be symmetric, i.e. give the same

---

[2]As O and T paragraphs are disjoint, this is the same as their concatenation.

result no matter whether $top_1 = O$ and $top_2 = T$ or vice-versa. Now consider another (extreme) case: lets say top_1 is half O and half T, with top_2 the same but with the other halves of O and T. In this case topic is not able to distinguish between O and T (beyond chance). What about cases in between the two extreme cases? Lets say, top_1 is 3/4 O and 1/4 T, and therefore top_2 is 1/4 O and 3/4 T (or vice versa - swap top_1 and top_2). In this case topics top_1 and top_2 are pretty good indicators of O and T, and a translationese classifier may pick up on topic signals rather than just translationese proper.

To design a measure that captures the relevance of topic classes to (binary) translationese classification, we need our measure to be symmetric, generalize to more than 2 topic classes, and factor in possible O and T class imbalance. To keep things simple[3], here we present a straightforward and easy to use measure we call *alignment of topic top_i with O and T*, denoted

$$align_{O,T}(top_i) \qquad (1)$$

with the majority class O or T covered by top_i (whatever it is for a given top_i) given the benefit of the doubt as the "correct" translationese class. We assume that $Data = O \cup T$, $O \cap T = \emptyset$, $\bigcup_{i=1}^{n} top_i = Data$ and $\bigcup_{i \neq j} top_i \cap top_j = \emptyset$, i.e. topic partitions our data, as does O and T. With this

$$align_{O,T}(top_i) = \frac{\max(|top_i \cap O|, |top_i \cap T|)}{|top_i|} \qquad (2)$$

$\max(\cdot, \cdot)$ makes the measure symmetric. Given $align_{O,T}(top_i)$, the weighted average is simply:

$$avg\_align_{O,T}(tops) = \sum_{i=1}^{n} w_i \times align_{O,T}(top_i) \qquad (3)$$

where a weight $w_i = |top_i|/|Data|$ is just the proportion of paragraphs in topic top_i divided by the total number of paragraphs in the data.

It is easy to see that the definition generalizes to n topic classes top_1 to top_n, that it adapts to different top_i topic sizes as well as the class imbalance between O and T[4]. $align_{O,T}(top_i) \in [0.5,$ 1] where $align_{O,T}(top_i) = 1$ signals perfect alignment of topic top_i with one of O or T, and that $align_{O,T}(top_i) = 0.5$ signals that top_i is maximally undecided with respect to O and T. And the same for $avg\_align_{O,T}(top)$.

Our alignment-based measure[5] is in fact the same as cluster *purity* (Zhao and Karypis, 2001) defined as

$$\frac{1}{M} \sum_{clu \in Cluster} \max_{cla \in Class} (clu \cap cla) \qquad (4)$$

where M is the size of the data, $Cluster$ and $Class$ the set of clusters and classes, respectively. With this we have

$$avg\_align_{Class}(Cluster)$$

$$= \sum_{clu \in Cluster} w_{clu} \times align_{Class}(clu)$$

$$= \sum_{clu \in Cluster} \frac{|clu|}{M} \times \frac{\max_{cla \in Class}(clu \cap cla)}{|clu|}$$

$$= \frac{1}{M} \sum_{clu \in Cluster} \max_{cla \in Class} (clu \cap cla) \qquad (5)$$

4.2 Experiments

We use LDA (Blei et al., 2001) as our main unsupervised topic model, as it provides a standard and well-understood baseline[6]. LDA makes two key assumptions: (1) documents are a mixture of topics, and (2) topics are a mixture of words. LDA generates a document-term matrix (DTM), where each document is represented by a row and the terms (words) corresponding to each document are represented by the columns. The DTM is decomposed into a document-topic matrix and a topic-word matrix. LDA assigns every word to a latent topic (top_i) through iteration, computing a topic word distribution (θ) in the data. To build this distribution, LDA uses two parameters: α which controls the per-document topic distribution, and β (the Dirichlet parameter) which controls the per-topic word distribution. LDA requires us to specify the number of topics n in advance. In our experiments we explore n over three orders of magnitude, roughly doubling n at each step, starting with $n = 2$ and going up to $n = 500$.

---

[3]We could design an entropy-based measure of the distribution of topic classes with respect to O and T, factor in classification probabilities of LDA etc.

[4]To see this, note that as O, T partition the data and as $\bigcup top_i$ also partition the data, if, let us say, $O \ll T$ and for some $top_i = O$, then there is nothing of O left to any of the other $top_{j \neq i}$. Note that in our data we have $|O| \approx |T|$.

[5]One of our reviewers suggested we compare our measure to existing cluster quality measures.

[6]We use the *Gensim* (Rehurek and Sojka, 2011) implementation of *Mallet* LDA (McCallum, 2002).

As LDA requires us to specify n in advance, we also use BERTopic (Grootendorst, 2022), which can find an optimal n given the data[7]. By default, BERTopic utilizes contextual sentence embeddings (SBERT), dimensionality reduction (UMAP), clustering (HDBSCAN), tokenizing (CountVectorizer), and a weighing scheme (c-TFIDF) to perform topic modeling. We choose the embedding model 'T-Systems-onsite/cross-en-de-roberta-sentence-transformer' from *Huggingface* (Wolf et al., 2020) and the defaults for all other modules.

For LDA, we explore a number of topics n with $n = 2, 5, 10, 20, 30, 50, 100, 200, 300, 400$ and 500, over three orders of magnitude. BERTopic returns 207 topics[8].

For each document (here paragraph), we use the highest probability LDA or BERTopic topic assigned to the document to label the document. A topic is then represented by the set of documents labeled with the topic. For each topic top_i, we compute how well the topic is aligned with O and T, i.e. we compute $align_{O,T}(top_i)$, and the weighted average over the topics: $avg\_align_{O,T}(top_1, ..., top_n)$.

4.3 Results

We plot $avg\_align_{O,T}(top_1, ..., top_n)$ in Fig. 1, varying n from 2 to 500, at each step roughly doubling n for LDA, and with $n = 207$ for and as determined by BERTopic. An average alignment of 0.5 (the dashed green line) shows topics maximally undecided with respect to O and T, while a score of 1 indicates perfect alignment where topics completely predict O and T. Fig. 1 shows topic alignment with O and T in the range of 0.55 to 0.62, depending on n, with BERTopic achieving the overall highest score of 0.62 at $n = 207$. For LDA, scores are highest (0.611 - 0.618) for $n = 10, 20,$ and 30. For good choices of topic numbers n, both LDA and BERTopic topics are able to predict O and T (i.e. translationese) by close to 0.62.

4.4 Discussion and Interpretation: the "Topic Floor" in Classification

This is an interesting and perhaps somewhat surprising result, but what exactly does it mean? There are two important caveats:

First, the fact that topic is able to predict O and T by close to 0.62 in the data does not necessarily mean (i.e. prove) that a high-performance BERT translationese classifier, such as the one presented in (Pylypenko et al., 2021), necessarily uses spurious topic information aligned with O and T in the data. At the same time, however, it cannot be ruled out. As a sanity check we tested how well a BERT classifier can learn to predict LDA topic classes for $n = 2, 10, 20$ and 30 and for BERTopic's 207. The results are $0.83, 0.64, 0.42, 0.44$ and 0.57 (all acc. and well above the largest class baseline). Given how well BERT-based classifiers can pick up patterns in the data, it is prudent to assume that BERT will be sensitive to and use topic signals spuriously aligned with O and T.

Second, we cannot at this stage completely rule out (other than perhaps through laborious manual inspection) that some LDA or BERTopic topics may in fact reflect genuine rather than spurious signals. LDA, e.g., uses lexical information, and perhaps some such information is a genuine translationese signal (unlike the place names clearly identified as a spurious topic signal in (Amponsah-Kaakyire et al., 2022)), such as e.g. certain forms of verbs (see Section 5 on the Known Unknowns).

The two caveats are aspects of the "unknown unknowns" we are chasing in this part of the paper. We have a clear indication that topic aligns with and hence can predict O and T in our data up to 0.62. There are very good reasons (but no proof) to assume that it is likely that a high-performance BERT classifier may use this, while we cannot completely rule out that some of the supposedly spurious topic signal may actually be genuine translationese. Given this, 0.62 is an *upper bound* of how well topic may predict O and T in our data. We may be well advised to take inspiration from the concept of a "noise floor" in Electronic Engineering. The noise floor is the hum and hiss (of a circuit) due to the components when there is no signal, and below which we cannot identify a signal. Given our findings, perhaps we should regard the 0.62 topic alignment with O and T in our data as a "topic floor" for translationese classification. This is in fact the recommendation we take from our work: instead of using 0.5 as a random baseline for our (roughly) balanced binary translationese data set, we should require 0.62 as established by the topic alignment experiments as a safe(r) baseline. Put differently, given our data we cannot really be
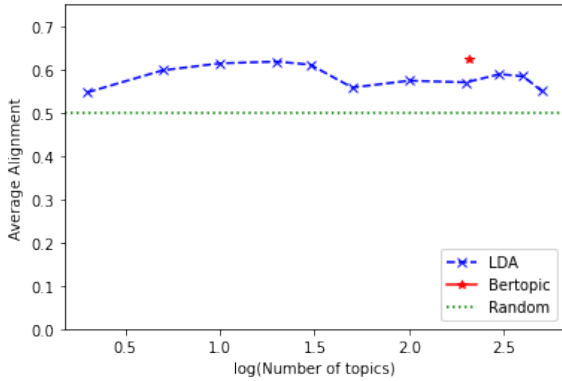
Figure 1: Average Topic - Target Classification Alignment $avg\_align_{O,T}(top_1, ..., top_n)$ for LDA and Bertopic

sure about O and T classification results ≤ 0.62. Suffice it to say, even with a 0.62 baseline, (most of) the classifiers presented in (Pylypenko et al., 2021; Amponsah-Kaakyire et al., 2022) easily surpass that baseline (with acc ≥ 0.9).

5 Chasing Known Unknowns

In this section, we assume that we have some knowledge about spurious topics in our data and that we want to both quantify and mitigate this spurious topic information in translationese classification. The difference in classification accuracy between a classifier that has access and one that does not have access to spurious topic information quantifies the impact of the spurious topic information in question. Using IG, Amponsah-Kaakyire et al. (2022) show that high-performance BERT-based classifiers use location names (Spanish names in the T, and German names in the O data) in the classification, clearly not a proper translationese but a spurious topic signal. Similarly, named entities (NEs) are often highly ranked in LDA topics. Therefore, the most straightforward approach towards mitigating specific spurious topic information in translationese classification is identifying NEs and masking them in the data.

5.1 Named Entity Recognition on Europarl

We focus on a scenario where we identify (and later mask) NEs automatically, rather than manually. While automatic NER is noisy, unlike potentially high-quality manual NER, it scales and constitutes a realistic application scenario. To assess NER performance, we experiment with a number of SOTA NER models, namely SpaCy (Honnibal and Montani, 2017), FLERT (Schweter and Akbik, 2020), multilingual-BERT (Devlin et al.,

NER model	Precision	Recall	F1-score
SpaCy	0.26	0.56	0.35
FLERT	0.39	0.35	0.37
mBERT-large	0.33	0.31	0.32
XLM-R-base	0.20	0.33	0.25
XLM-R-large	0.19	0.31	0.24
DistilBERT-large	0.34	0.31	0.32
BERT-German	0.65	0.42	0.52

Table 1: Comparison table for NER models on the gold standard dataset from (Agerri et al., 2018)

2019b), XLM-RoBERTa (Conneau et al., 2020), DistilBERT (Sanh et al., 2019), and BERT-German (Chan et al., 2020), comparing F1, precision and recall for each of the models against a gold standard NE tagged dataset (Agerri et al., 2018). The gold standard consists of 800 sentences from the Europarl German data manually annotated following the ConLL 2002 (Tjong Kim Sang, 2002) and 2003 (Tjong Kim Sang and De Meulder, 2003) guidelines, with a total of 433 named entities.

Table 1 shows that BERT-German (a fine-tuned version of bert-base-multilingual-cased on the German WikiANN dataset) has the highest precision, second-highest recall and the highest F1 (0.52) among all the NER models. Hence, we choose BERT-German for all our NER experiments.

5.2 Translationese Classification

To quantify the impact of NE-based spurious topic information on translationese classification, we modify our data by masking NEs (Section 5.2.1), and, in a separate experiment, we explore full masking of the data using POS (Section 5.2.2).

Following (Pylypenko et al., 2021), we use multilingual BERT (Devlin et al., 2019b) (BERT-base-multilingual-uncased) and fine-tune BERT on the training set of our data set using the *Huggingface* library. We use a batch size of 16, a learning rate of $4 \cdot 10^{-5}$, and the Adam optimizer with $\epsilon = 1 \cdot 10^{-8}$.

We compare our models with the BERT model reproduced from (Pylypenko et al., 2021): a pretrained BERT-base model (12 layers, 768 hidden dimensions, 12 attention heads) fine-tuned on translationese classification using unmasked data.

5.2.1 Named Entity Masking

We use Bert-German to replace NEs with one of three course-grained NE-type tags: [LOC], [PER], and [ORG]. For example, the unmasked string "*John will go to Berlin.*" is NE masked as "[PER] *will go to* [LOC]". In our train-dev-test sets, we have 202036, 42072, and 43489 NEs.

We carry out experiments with four train-test configurations: masked-masked, unmasked-masked, masked-unmasked, and the original unmasked-unmasked. For each of these configurations, we fine-tune BERT to the specifics of the training set (i.e. masked or unmasked). We supply the three NE-type tags as special additional tokens to the BERT tokenizer to ensure that they are consistently represented by their NE-type token (and that no sub-word splitting is applied to NE-type tokens). If NEs are responsible for spurious topic information in translationese classification, we expect masking NEs to mitigate spurious correlations in the data and to result in reduced translationese classification accuracies, allowing us to quantify this aspect of spurious topic correlations.

5.2.2 Part-Of-Speech (POS) Full Masking

To analyze BERT's performance on fully delexicalized data, we use the finer POS tagger from SpaCy (Honnibal and Montani, 2017) which utilizes the TIGER Treebank (Brants et al., 2004)[9]. Given: "*Jetzt solle erneut ein Antrag gestellt werden .*", the POS tag sequence is: "ADV VMFIN ADJD ART NN VVPP VAINF $.". We pretrain BERT on POS-tagged data from 3% of the German Wikipedia dump (1.5 million sentences) on the BertforMaskedLM objective for 2 epochs. We use BertWordPieceTokenizer for tokenization. To adjust the BERT model to the small vocabulary, we use only 6 encoder layers (instead of 12), a learning rate of 5.10^{-6}, and the Adam optimizer with $\epsilon = 1.10^{-8}$. We use the POS-pretrained model and fine-tune it with the POS-tagged monolingual German dataset from the MPDE corpus (Amponsah-Kaakyire et al., 2021) (same fine-tuning parameters as other experiments).

5.3 Integrated Gradients (IG)

Amponsah-Kaakyire et al. (2022) use IG attribution scores to show that BERT utilizes spurious correlations in the data, for example, German T data translated from Spanish contain mentions of Spanish geographical areas, such as 'Spanien', 'Barcelona' etc. as top tokens identified by IG. Here we use IG on BERT trained on masked data, and compute the top tokens with the highest attribution scores on average across the masked test sets. We also compute the top POS tags by performing IG on

---

[9] https://github.com/explosion/spaCy/blob/master/spacy/glossary.py (last accessed 11 Aug, 2023)

BERT trained on fully POS-tagged data.

5.4 Results

Train-Test	Test Set Acc (%)	95% CI
m-m	0.89±0.00	[0.88,0.89]
m-u	0.89±0.00	[0.89,0.90]
u-m	0.90±0.00	[0.90,0.91]
u-u	0.92±0.00	[0.91,0.93]

Table 2: NE masked experiments pretrained-BERT-ft Acc(uracy); CI(Conf. Interval); m(asked), u(nmasked).

Test Set Acc (%)	95% CI
0.78 ± 0.00	[0.77, 0.79]

Table 3: POS-masked experiments POS BERT fine-tuned with TIGER Treebank tags, Acc(uracy); CI(Conf. Interval)

	Translationese		Original	
	Token	AAS	Token	AAS
1	besuchte	0.61	•	0.83
2	entdeckte	0.60	alpen	0.69
3	veroffentlichte	0.53	apo	0.66
4	gehorten	0.51	profits	0.63
5	fuhrte	0.47	##nova	0.59
6	nominal	0.46	super	0.49
7	benutzt	0.46	##bud	0.48
8	tari	0.45	##ndus	0.46
9	starb	0.44	##enland	0.46
10	eman	0.43	##hutte	0.45
11	loste	0.39	digitale	0.45
12	planeten	0.39	ros	0.45
13	geboren	0.38	population	0.43
14	veroffentlichten	0.38	pla	0.43
15	neige	0.37	express	0.42
16	schrieb	0.37	##vagen	0.40
17	priester	0.36	stahl	0.40
18	scheiterte	0.36	ez	0.40
19	genus	0.35	stands	0.40
20	territorium	0.35	##nog	0.39

Table 4: Top-20 tokens with highest IG average attribution score (AAS) for the NE-masked test set.

5.4.1 Results NE Masking

Table 2 shows test set accuracies for the NE masking experiments outlined in Section 5.2.1. We use Bootstrap Resampling, with 100 samples and 95% confidence intervals. Results (u-u against all others) are statistically significant. Consistent with expectation, under all training-test data conditions, masking NE-related information lowers classification results. Compared to the (Amponsah-Kaakyire et al., 2021) unmasked-unmasked baseline, the performance drop is between 0.026-0.032 points absolute. In absolute terms, the performance drop incurred in mitigating spurious topic information

	Translationese		Original	
	Token	AAS	Token	AAS
1	APPO (ADP)	0.32	ADV (ADV)	0.21
2	PRELS (PRON)	0.19	. (PUNCT)	0.12
3	KOUI (SCONJ)	0.15	TRUNC (X)	0.06
4	PPOSAT (DET)	0.14	ADJD (ADJ)	0.05
5	PRELAT (DET)	0.12	FM (X)	0.04
6	PIS (PRON)	0.12	PROAV (ADV)	0.03
7	PPER (PRON)	0.11	PDS (PRON)	0.03
8	PDAT (DET)	0.11	VVIZU (VERB)	0.02
9	VVFIN (VERB)	0.11	PTKANT (PART)	0.02
10	VMFIN (VERB)	0.10	PTKZU (PART)	0.01

Table 5: Top-10 tokens with highest IG average attribution score (AAS) for the POS-tagged test set (TIGER Treebank tags). Corresponding UPOS tags are given in braces. We use the conversion table from `https://universaldependencies.org/tagset-conversion/de-stts-uposf.html` (last accessed 11 Aug, 2023).

in terms of NEs masking is visible but small, in the order of 3 to 4 % points absolute, if classification accuracy is expressed as % points. This indicates that this type of spurious topic information and the ensuing "Clever Hans" is a small part of the strong BERT translationese classification performance.

5.4.2 Results Full POS Masking

Table 3 shows that translationese classification results on fully de-lexicalized POS-masked data are much lower than for NE-masked data[10]. In this regime, BERT is missing much valuable information. At the same time, the classification accuracy of almost 0.78 shows that BERT is able to pick up on morpho-syntactic aspects of translationese. We also performed an analogous experiment with Universal POS tags (see Appendix, section A.3), and obtained accuracy of almost 0.77.

5.4.3 Integrated Gradients NE and POS

Table 4 shows the top-20 IG token attributions for O and T data in the masked-masked condition. Unlike (Amponsah-Kaakyire et al., 2022) the "translationese" column does not show any Spanish or other place names. At the same time the "original" column still contains a few location tokens (or possible subwords of location tokens), such as "alpen", "##enland", "ez" (as in (Amponsah-Kaakyire et al., 2022)), confirming the fact that automatic NER is not perfect (see P, R and F1 scores in Table 1).

It is interesting to note that hand-in-hand with the reduction of spurious NE-based cues, many of the observations from (Amponsah-Kaakyire et al.,

---

[10]For fully PoS-masked data it does not make sense to report mixed train-test conditions.

2022) are confirmed and in fact come to the fore. In the T class we observe an even stronger presence of verbs in Präteritum form (this time not only regular, but also irregular verbs), which Amponsah-Kaakyire et al. (2022) link to the fact that translators might have preferred to use a more written style while translating the transcribed speeches.

Table 5 presents the top-10 IG attributed tokens for the POS-tagged test set, for BERT trained on POS-tagged data. Interestingly, the top tags are APPO (postpositions) for the translationese class T, and ADV (adverbs) for the originals class O, which confirms findings in (Pylypenko et al., 2021) who show that relative frequencies of adverbs and adpositions are among the highest-ranked features that correlate with predictions of various translationese classification architectures, including BERT (even though their experiment is performed in the multilingual setting, and not on just German data like ours). They also show that the ratio of determiners is an important feature, and we see many tags corresponding to this category in our list: PPOSAT (attributive possessive pronouns), PRELAT (attributive relative pronouns), PDAT (attributive demonstrative pronouns), etc. The results for UPOS tags are similar (see Appendix, section A.3).

6 Conclusion

We present a measure that, given a data set and target classification labels, quantifies the possible impact of *unknown spurious topic information on classification*. The measure is based on aligning unsupervised topics with target labels and is equivalent to *purity* in clustering. We propose the concept of a "topic floor" (akin to "noise floor") as an upper bound of the impact of spurious topic information on classification in classification. We use masking to quantify and mitigate *known spurious topic information*. We present empirical results for topic floor and masking to quantify "Clever Hans" in the translationese data of (Amponsah-Kaakyire et al., 2022). We use IG attribution to show that in masked settings where known spurious correlations are mitigated, BERT learns features closer to proper translationese.

References

Rodrigo Agerri, Yiling Chung, Itziar Aldabe, Nora Aranberri, Gorka Labaka, and German Rigau. 2018. Building named entity recognition taggers via parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kwabena Amponsah-Kaakyire, Daria Pylypenko, Cristina España-Bonet, and Josef van Genabith. 2021. Do not rely on relay translations: Multilingual parallel direct Europarl. In *Proceedings for the First Workshop on Modelling Translation: Translatology in the Digital Age*, pages 1–7, online. Association for Computational Linguistics.

Kwabena Amponsah-Kaakyire, Daria Pylypenko, Josef Genabith, and Cristina España-Bonet. 2022. Explaining translationese: why are neural classifiers better and what do they learn? In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 281–296, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Ehud Alexander Avner, Noam Ordan, and Shuly Wintner. 2016. Identifying translationese at the word and sub-word level. *Digit. Scholarsh. Humanit.*, 31:30–54.

Mona Baker, Gill Francis, and Elena Tognini-Bonelli. 1993. *Text and technology: in honour of John Sinclair*. John Benjamins Publishing.

Marco Baroni and Silvia Bernardini. 2005. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. *Literary and Linguistic Computing*, 21(3):259–274.

David Blei, Andrew Ng, and Michael Jordan. 2001. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.

Caroline Bogaert. 2011. *Is absolute multilingualism maintainable? The language policy of the European Parliament and the threat of English as a lingua franca*. Ph.D. thesis, MA thesis, Universiteit Gent.

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. Tiger: Linguistic interpretation of a german corpus. *Research on language and computation*, 2:597–620.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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.

Koel Dutta Chowdhury, Rricha Jalota, Cristina España-Bonet, and Josef Genabith. 2022. Towards debiasing translation artifacts. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3983–3991, Seattle, United States. Association for Computational Linguistics.

Markus Freitag, Isaac Caswell, and Scott Roy. 2019. APE at scale and its implications on MT evaluation biases. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 34–44, Florence, Italy. Association for Computational Linguistics.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

José Hernández-Orallo. 2019. Gazing into clever hans machines. *Nature Machine Intelligence*, 1(4):172–173.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Iustina Ilisei, Diana Inkpen, Gloria Corpas Pastor, and Ruslan Mitkov. 2010. Identification of translationese: A machine learning approach. In *International conference on intelligent text processing and computational linguistics*, pages 503–511. Springer.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8.

Andrew Kachites McCallum. 2002. Mallet: A machine learning for language toolkit. Http://mallet.cs.umass.edu.

Jingwei Ni, Zhijing Jin, Markus Freitag, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Original or translated? a causal analysis of the impact of translationese on machine translation performance. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5303–5320, Seattle, United States. Association for Computational Linguistics.

Tiina Puurtinen. 2003. Genre-specific features of translationese? linguistic differences between translated and non-translated finnish children's literature. *Literary and linguistic computing*, 18(4):389–406.

Daria Pylypenko, Kwabena Amponsah-Kaakyire, Koel Dutta Chowdhury, Josef van Genabith, and Cristina España-Bonet. 2021. Comparing feature-engineering and feature-learning approaches for multilingual translationese classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8596–8611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ella Rabinovich and Shuly Wintner. 2015. Unsupervised Identification of Translationese. *Transactions of the Association for Computational Linguistics*, 3:419–432.

Radim Rehurek and Petr Sojka. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. 2016. Information density and quality estimation features as translationese indicators for human translation classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 960–970, San Diego, California. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Stefan Schweter and Alan Akbik. 2020. Flert: Document-level features for named entity recognition.

Jasdeep Singh, Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2019. Xlda: Cross-lingual data augmentation for natural language inference and question answering.

Ilia Sominsky and Shuly Wintner. 2019. Automatic detection of translation direction. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1131–1140, Varna, Bulgaria. INCOMA Ltd.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Elke Teich. 2012. *Cross-Linguistic Variation in System and Text*. De Gruyter Mouton, Berlin, Boston.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Gideon Toury. 1980. *In search of a theory of translation*. Porter Institute for Poetics and Semiotics, Tel Aviv University.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

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. Trans-formers: 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.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

Ying Zhao and George Karypis. 2001. *Criterion functions for document clustering: Experiments and analysis*. Retrieved from the University of Minnesota Digital Conservancy.

A   Appendices

A.1   Topics by LDA and Bertopic

Here we take a closer look at LDA and Bertopic topics. While we do not find much evidence for geographic LDA topics, we do find quite a few geographic BERTtopic topics, for example, Topic 10 consists of word tokens "türkei, türkischen, türkische, kriterien, helsinki, daß, politischen, kurdischen, menschenrechte, die", Topic 14: "palästinensischen, israel, arafat, israelischen, palästinensische, sharon, autonomiebehörde, palästinenser, frieden, israels", Topic 23: "kuba, kubanischen, kubaner, kubas, kubanische, dissidenten, volk, castro, cotonou, havanna" predominantly consist of geographical terms.

A.2   Topic Classification

To understand if BERT is able to learn the topics identified by the topic modeling experiments, we perform topic classification by finetuning pre-trained-BERT on the topics found by LDA and BERTopic.

We use a similar ratio for each topic as the train:dev:test (29580:6336:6344) ratio for the translationse classification experiments.

n	Test Set Acc (%)	95% CI	Baseline Acc
2	0.832±0.00	[0.83,0.84]	0.50
10	0.636±0.01	[0.62,0.64]	0.18
20	0.417±0.00	[0.41,0.42]	0.13
30	0.442±0.00	[0.44,0.45]	0.11
207 (BT)	0.569±0.00	[0.56,0.57]	0.002

Table 6: Topic Classification experiments pretrained-BERT-ft Acc(uracy); n(umber of topics), CI(Conf. Interval), BT (BERTopic).

Table 6 shows the topic classification results for the topics output by LDA and BERTopic. We also show baseline accuracies, when the model only predicts the largest class.

A.3   Full UPOS Masking

Apart from full masking with detailed tags from the TIGER Treebank, we also explore full masking with the more general Universal POS tags. BERT was pre-trained and fine-tuned on the POS-tagged data in the same way as described in section 5.2.2 for the TIGER tags. The translationese classification accuracy (Table 7) is slightly lower than for the detailed tags (Table 3).

Test Set Acc (%)	95% CI
0.768 ± 0.00	[0.76, 0.77]

Table 7: POS-masked experiments POS BERT fine-tuned with UPOS tags, Acc(uracy); CI(Conf. Interval)

IG results (Table 8) show patterns similar to those for the detailed tags (Table 5): PRON, DET and SCONJ for translationese; X, ADV, PART and ADJ for originals.

	Translationese		Original	
	Token	AAS	Token	AAS
1	PRON	0.12	X	0.31
2	PUNCT	0.08	ADV	0.16
3	CCONJ	0.07	PART	0.07
4	DET	0.06	AUX	0.05
5	SCONJ	0.04	VERB	0.03
6	NOUN	0.02	NUM	0.02
7			PROPN	0.02
8			ADJ	0.02
9			NOUN	0.02

Table 8: Top-10 tokens with highest IG average attribution score (AAS) for the POS-tagged test set (UPOS tags).