Natural Language Engineering (2023), **29**, pp. 554–583
doi:10.1017/S1351324922000225

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Ad astra or astray: Exploring linguistic knowledge of multilingual BERT through NLI task

Maria Tikhonova[1,*], Vladislav Mikhailov[1], Dina Pisarevskaya[2], Valentin Malykh[3] and Tatiana Shavrina[1,4,*]

[1]HSE University, Moscow, Russia, [2]Independent Reserecher, London, UK, [3]Huawei Noah's Ark Lab, Moscow, Russia, and [4]AI Research Institute (AIRI), Moscow, Russia
*Corresponding author. E-mail: m_tikhonova94@mail.ru; rybolos@gmail.com

(Received 7 June 2021; revised 27 April 2022; accepted 2 May 2022; first published online 9 June 2022)

Abstract

Recent research has reported that standard fine-tuning approaches can be *unstable* due to being prone to various sources of randomness, including but not limited to weight initialization, training data order, and hardware. Such brittleness can lead to different evaluation results, prediction confidences, and generalization inconsistency of the same models independently fine-tuned under the same experimental setup. Our paper explores this problem in natural language inference, a common task in benchmarking practices, and extends the ongoing research to the multilingual setting. We propose six novel textual entailment and broad-coverage diagnostic datasets for French, German, and Swedish. Our key findings are that the mBERT model demonstrates fine-tuning instability for categories that involve lexical semantics, logic, and predicate-argument structure and struggles to learn monotonicity, negation, numeracy, and symmetry. We also observe that using extra training data only in English can enhance the generalization performance and fine-tuning stability, which we attribute to the cross-lingual transfer capabilities. However, the ratio of particular features in the additional training data might rather hurt the performance for model instances. We are publicly releasing the datasets, hoping to foster the diagnostic investigation of language models (LMs) in a cross-lingual scenario, particularly in terms of benchmarking, which might promote a more holistic understanding of multilingualism in LMs and cross-lingual knowledge transfer.

Keywords: Evaluation; Model Interpretation; Multilinguality; Natural Language Inference; Cross-lingual learning; Transfer learning

1. Introduction

The latest advances in neural architectures of language models (LMs) (Vaswani et al., 2017) have raised the importance of NLU benchmarks as a standardized practice of tracking progress in the field and exceeded conservative human baselines on some datasets (Raffel et al., 2020; He et al., 2021). Such LMs are centered around the "pre-train & fine-tune" paradigm, where a pre-trained LM is directly fine-tuned for solving a downstream task. Despite the impressive empirical results, pretrained LMs struggle to learn linguistic phenomena from raw text corpora (Rogers 2021), even when increasing the size of pretraining data (Zhang et al., 2021). Furthermore, the fine-tuning procedure can be unstable (Devlin et al., 2019) and raise doubts about whether it promotes task-specific linguistic reasoning (Kovaleva et al., 2019). The brittleness of standard fine-tuning approaches to various sources of randomness (e.g., weight initialization and training data order) can lead to different evaluation results and prediction confidences of models,

Article last updated 5th April 2023.

independently fine-tuned under the same experimental setup. Recent research has defined this problem as *(in)stability* (Dodge et al., 2020); (Mosbach et al., 2020a), which now serves as a subject of an interpretation direction, aimed at exploring the consistency of linguistic generalization of LMs (McCoy et al., 2018, 2020).

Our paper is devoted to this problem in the task of natural language inference (NLI) which has been widely used to assess language understanding capabilities of LMs in monolingual and multilingual benchmarks (Wang et al., 2018, 2019; Liang et al., 2020; Hu et al., 2020b). The task is framed as a binary classification problem, where the model should predict if the meaning of the *hypothesis* is entailed with the *premise*. Many works show that NLI models learn shallow heuristics and spurious correlations in the training data (Naik et al., 2018; Glockner et al., 2018; Sanchez et al., 2018), stimulating a targeted evaluation of LMs on out-of-distribution sets covering inference phenomena of interest (Yanaka et al., 2019b; Yanaka et al., 2019a; McCoy et al., 2019; Tanchip et al., 2020). Although such datasets are extremely useful for analyzing how well LMs capture inference and abstract properties of language, English remains the focal point of the research, leaving other languages underexplored.

To this end, our work extends the ongoing research on the fine-tuning stability and consistency of linguistic generalization to the multilingual setting, covering five Indo-European languages from four language groups: English (West Germanic), Russian (Balto-Slavic), French (Romance), German (West Germanic), and Swedish (North Germanic). The contributions are summarized as twofold. First, we propose GLUE-style textual entailment and diagnostic datasets[a] for French, Swedish, and German. Second, we explore the *stability* of linguistic generalization of mBERT across five languages mentioned above, analyzing the impact of the random seed choice, training dataset size, and presence of linguistic categories in the training data. Our work differs from similar approaches described in Section 2 in that we (i) evaluate the inference abilities through the lens of broad-coverage diagnostics, which is often neglected for upcoming LMs, typically compared among one another only by the averaged scores on canonical benchmarks (Dehghani et al., 2021); and (ii) analyze the *per-category* stability of the model fine-tuning for the considered languages, testing mBERT's cross-lingual transfer abilities.

2. Related work

NLI and diagnostic datasets. There is a wide variety of datasets constructed to facilitate the development of novel approaches to the problem of NLI (Storks et al., 2019). The task has evolved within a series of RTE challenges (Dagan et al., 2005) and now comprises several standardized benchmark datasets such as SICK (Marelli et al., 2014), SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), and XNLI (Conneau et al., 2018b). Despite the rapid progress, recent work has found that these benchmarks may contain biases and annotation artifacts which raise questions whether state-of-the-art models indeed have or acquire the inference abilities (Tsuchiya 2018; Belinkov et al., 2019). Various linguistic datasets have been proposed to challenge the models and help to improve their performance on inference features (Glockner et al., 2018; Yanaka et al., 2019a, 2019b, 2020; McCoy et al., 2019; Richardson et al., 2020; Hossain et al., 2020; Tanchip et al., 2020). The MED (Yanaka et al., 2019a) and HELP (Yanaka et al., 2019b) datasets focus on aspects of monotonicity reasoning, motivating the follow-up work on systematicity of this phenomenon (Yanaka et al., 2020). HANS (McCoy et al., 2019) aims at evaluating the generalization abilities of NLI models beyond memorizing lexical and syntactic heuristics in the training data. Similar in spirit, the concept of semantic fragments has been applied to synthesize datasets that target quantifiers, conditionals, monotonicity reasoning, and other features (Richardson et al., 2020). The SIS dataset (Tanchip et al., 2020) covers symmetry of verb predicates, and it is designed to improve systematicity in neural models. Another feature studied in the field is negation which has proved

[a] https://github.com/MariyaTikhonova/multilingual_diagnostics/

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

to be challenging not only for the NLI task (Hossain et al., 2020; Hosseini et al., 2021) but also for probing factual knowledge in masked LMs (Kassner and Schütze 2020).

Last but not least, broad-coverage diagnostics is introduced in the GLUE benchmark (Wang et al., 2018) and has now become a standard dataset for examining linguistic knowledge of LMs on GLUE-style leaderboards. To the best of our knowledge, there are only two counterparts of the diagnostic dataset for Chinese and Russian, introduced in the CLUE (Xu et al., 2020) and Russian SuperGLUE benchmarks (Shavrina et al., 2020). Creating such datasets is not addressed in recently proposed GLUE-like benchmarks for Polish (Rybak et al., 2020) and French (Le et al., 2020).

Stability of neural models. A growing body of recent studies has explored the role of optimization, data, and implementation choices on the stability of training and fine-tuning neural models (Henderson et al., 2018; Madhyastha and Jain 2019; Dodge et al., 2020; Mosbach et al., 2020a). Bhojanapalli et al., (2021) and Zhuang et al., (2021) investigate the impact of weight initialization, mini-batch ordering, data augmentation, and hardware on the prediction disagreement between image classification models. In NLP, BERT has demonstrated instability when being fine-tuned on small datasets across multiple restarts (Devlin et al., 2019). This has motivated further research on the most contributing factors to such behavior, mostly the dataset size and the choice of random seed as a hyperparameter (Bengio 2012), which influences training data order and weight initialization. The studies report that changing *only* random seed during the fine-tuning stage can cause a significant standard deviation of the validation performance, including tasks from the GLUE benchmark (Lee et al., 2019; Dodge et al., 2020; Mosbach et al., 2020a; Hua et al., 2021). Another direction involves studying the effect of random seeds on model performance and robustness in terms of attention interpretation and gradient-based feature importance methods (Madhyastha and Jain 2019).

Linguistic competence of BERT. A plethora of works is devoted to the linguistic analysis of BERT, and the inspection of how fine-tuning affects the model knowledge (Rogers et al., 2020). The research has covered various linguistic phenomena, including syntactic properties (Warstadt and Bowman 2019), structural information (Jawahar et al., 2019), semantic knowledge (Goldberg 2019), common sense (Cui et al., 2020), and many others (Ettinger 2020). Contrary to the common understanding that BERT can capture the language properties, some studies reveal that the model tends to lose the information after fine-tuning (Miaschi et al., 2020); (Singh et al., 2020); (Mosbach et al., 2020b) and fails to acquire task-specific linguistic reasoning (Kovaleva et al., 2019); (Zhao and Bethard 2020); (Merchant et al., 2020). Several works explore the consistency of linguistic generalization of neural models by independently training them from 50 to 5,000 times and evaluating their generalization performance (Weber et al., 2018; Liška et al., 2018; McCoy et al., 2018; McCoy et al., 2020). In the spirit of these studies, we analyze the stability of the mBERT model w.r.t. diagnostic inference features, extending the experimental setup to the multilingual setting.

3. Multilingual datasets

This section describes textual entailment and diagnostic datasets for five Indo-European languages: English (West Germanic), Russian (Balto-Slavic), French (Romance), German (West Germanic), and Swedish (North Germanic). We use existing datasets for English (Wang et al., 2019) and Russian (Shavrina et al., 2020) and propose their counterparts for the other languages based on the GLUE-style methodology (Wang et al., 2018).

3.1 Recognizing textual entailment

The task of recognizing textual entailment is framed as a binary classification problem, where the model should predict if the meaning of the *hypothesis* is entailed with the *premise*. We provide an example from the English RTE dataset below and describe brief statistics for each language in Table 1.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

Table 1. Statistics of the NLI datasets. **Vocab size** refers to the total number of unique words. **Num. of words** stands for the average number of words in a sample. **Fr** = French; **De** = German; **Sw** = Swedish.

Task	Train	Validation	Test	Vocab size	Num. of words
RTE	2490	277	3000	22,200	26.9
TERRa	2616	307	3198	23,300	19.5
TERRa (Fr)	2616	307	3198	13,300	27.5
TERRa (De)	2616	306	3197	17,100	24.1
TERRa (Sw)	2613	307	3194	14,500	21.3

- Premise: *'Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation.'*
- Hypothesis: *'Christopher Reeve had an accident.'*
- Entailment: **False**.

English: *RTE* (Wang et al., 2018) is a collection of datasets from a series of competitions on recognizing textual entailment, constructed from news and Wikipedia (Dagan et al., 2005; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009).

Russian: *Textual Entailment Recognition for Russian (TERRa)* (Shavrina et al., 2020) is an analog of the RTE dataset that consists of sentence pairs sampled from news and fiction segments of the Taiga corpus (Shavrina and Shapovalova 2017).

French, German, Swedish: Each sample from *TERRa* is manually translated and verified by professional translators with the linguistic peculiarities preserved, culture-specific elements localized, and ambiguous samples filtered out. The resulting datasets contain fewer unique words than the ones constructed by filtering text sources (*RTE* and *TERRa*). We relate this to the fact that translated texts may exhibit less lexical diversity and vocabulary richness (Al-Shabab 1996; Nisioi et al., 2016).

3.2 Broad-coverage diagnostics

Broad-coverage diagnostics (Wang et al., 2018) is an expert-constructed evaluation dataset that consists of 1104 NLI sentence pairs annotated with linguistic phenomena under four high-level categories (see Table 2). The dataset is originally included in the GLUE benchmark. It is used as an additional test set for examining the linguistic competence of LMs, which allows for revealing possible biases and conducting a systematic analysis of the model behavior.

As part of this study, *LiDiRus* (Linguistic Diagnostics for Russian), an equivalent diagnostic dataset for the Russian language, is created (Shavrina et al., 2020). The creation procedure includes a manual translation of the English diagnostic samples by expert linguists so that each indicated linguistic phenomenon and target label is preserved and culture-specific elements are localized. We apply the same procedure to construct diagnostic datasets for French, German, and Swedish by translating and localizing the English diagnostic samples. The label distribution in each dataset is 42/58% (Entailment: **True**/**False**). Consider an example of the NLI pair (Sentence 1: *'John married Gary'*; Sentence 2: *'Gary married John'*; Entailment: **True**) and its translation in each language:

- **English**: *'John married Gary'* entails *'Gary married John'*;
- **Russian**: ' *Боб женился на Алисе*' entails ' *Алиса вышла замуж за Боба*';

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

Table 2. The linguistic annotation of the diagnostic dataset.

High-level categories	Low-level categories
Lexical semantics	Lexical entailment, morphological negation, factivity, symmetry/collectivity, redundancy, named entities, quantifiers
Predicate-argument structure	Core arguments, prepositional phrases, ellipsis/implicits, anaphora/coreference, active/passive, nominalization, genitives/partitives, datives, relative clauses, coordination scope, intersectivity, restrictivity
Logic	Negation, double negation, intervals/numbers, conjunction, disjunction, conditionals, universal, existential, temporal, upward monotone, downward monotone, non-monotone
Knowledge	Common sense, world knowledge

- **French**: *'John a épousé* entails *'Gary a épousé John'*;
- **German**: *'John heiratete Gary'* entails *'Gary heiratete John'*;
- **Swedish**: *'John gifte sig med Gary'* entails *'Gary gifte sig med John'*.

Linguistic challenges. Special attention is paid to the problems of the feature-wise translation of the examples. Since the considered languages are Indo-European, there appear fewer translation challenges. For instance, all languages have morphological negation mechanisms, lexical semantics features, common sense, and world knowledge instances. The main distinctions are related to the category of the *Predicate-Argument Structure*. The strategy of case coding is exhibited differently across the languages, for example, in dative constructions. Dative was widely used in all ancient Indo-European languages and is still present in modern Russian, retaining numerous functions. In contrast, dative constructions are primarily underrepresented in English and Swedish, and all the dative examples in the translations involve impersonal constructions with an indirect object instead of a subject. The same goes for genitives and partitives, where standard noun phrase syntax indicates genitive relations as Swedish and English do not have case marking. For French, the *"de + noun"* constructions are used to indicate partitiveness or genitiveness. Below is an example of an English sentence and its corresponding translations to Swedish and French:

- **English:** *'A formation of approximately 50 officers of the police of the City of Baltimore eventually placed themselves between the rioters and the militiamen, allowing the 6th Massachusetts to proceed to Camden Station.'*;
- **Swedish:** *'Om 50 poliser i staden Baltimore, i slutändan stod mellan demonstranterna och brottsbekämpande myndigheter, vilket gjorde det möjligt för 6: e Massachusetts Volunteer Regiment går till Cadman station.'*;
- **French:** *'Une cinquantaine de policiers de Baltimore se sont finalement interposés entre les manifestants et les forces de l'ordre, permettant au 6e régiment de volontaires du Massachusetts de se rendre à Cadman Station.'*.

Translations for the *Logic* and *Knowledge* categories are obtained with no difficulty, for example, all *existential* constructions share patterns with the translated analogs of the quantifiers such as *"some," "many,"* etc. However, we acknowledge that some low-level categories cannot be forwardly translated. For example, elliptic structures, are in general, quite different in Russian than in the other languages. Despite this, the translation-based method avoids the need for additional language-specific expert annotation.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

4. Experimental setup

The experiments are conducted on the mBERT[b] model, pretrained on concatenated monolingual Wikipedia corpora in 104 languages. We use the SuperGLUE framework under the jiant environment (Pruksachatkun et al., 2020b) to fine-tune the model multiple times for each language with a fixed set of hyperparameters while changing *only* the random seeds.

Fine-tuning. We follow the SuperGLUE fine-tuning and evaluation strategy with a set of default hyperparameters as follows. We fine-tune the mBERT model using a random seed $\in [0;5]$, batch size of 4, learning rate of $1e^{-5}$, global gradient clipping, dropout probability of $p = 0.1$, and the *AdamW* optimizer (Loshchilov and Hutter 2017). The fine-tuning is performed on 4 Christofari[c] Tesla V100 GPUs (32GB) for the maximum number of 10 epochs with early stopping on the NLI validation data. The model is evaluated on the corresponding broad-coverage diagnostics dataset as described below.

Evaluation. Since the feature distribution and class ratio in the diagnostic set are not balanced, the model performance is evaluated with Matthew's correlation coefficient (MCC), the two-class variant of the R_3 metric (Gorodkin 2004):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

MCC is computed between the array of model predictions and the array of gold labels (Entailment: **True**/**False**) for each low-level linguistic feature according to the annotation (Wang et al., 2019). The range of values is $[-1;1]$ (higher is better).

Fine-tuning stability. Fine-tuning stability has multiple definitions in recent research. The majority of studies estimate the stability as the *standard deviation* of the validation performance, measured by accuracy, MCC, or F1-score (Phang et al., 2018; Lee et al., 2019; Dodge et al., 2020). Another possible notion is *per-point stability*, where a set of models is analyzed w.r.t. their predictions on the same evaluation sample (Mosbach et al., 2020a; McCoy et al., 2019). More recent works evaluate the stability by more granular measures, such as predictive divergence, L2 norm of the trained weights, and standard deviation of subgroup validation performance (Zhuang et al., 2021). This work analyzes the stability in terms of pairwise Pearson's correlation as follows. Given a fixed experimental setup, we compute the correlation coefficients between the MCC scores on the diagnostic datasets, achieved by the models trained with different random seeds, and average the coefficients by the total number of models (higher is better). Besides, we assess the *per-category* stability, that is, the standard deviation in the model performance w.r.t. random seeds for samples within a particular diagnostic category.

5. Testing the linguistic knowledge and fine-tuning stability

5.1 Language-wise diagnostics

We start with investigating how well the linguistic properties are learned given the standardized NLI dataset by fine-tuning the mBERT model on the corresponding train data for each language independently with the same hyperparameters and computing *overall MCC* by averaging MCC scores for each diagnostic feature. Figure 1 shows a language-wise heat map with the results we use as a "baseline" performance to analyze different experiment settings. Despite the fact that the overall MCC scores are insignificantly different from one another (e.g., **German**: 0.15, **English**: 0.2), there is variability in how the model outputs correlate with the linguistic features w.r.t. the languages. In order to measure this variability, we compute pairwise Pearson's correlation between the overall MCC scores and average the coefficients over the total number of language pairs. The

[b] https://huggingface.co/bert-base-multilingual-cased
[c] https://sbercloud.ru/en

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

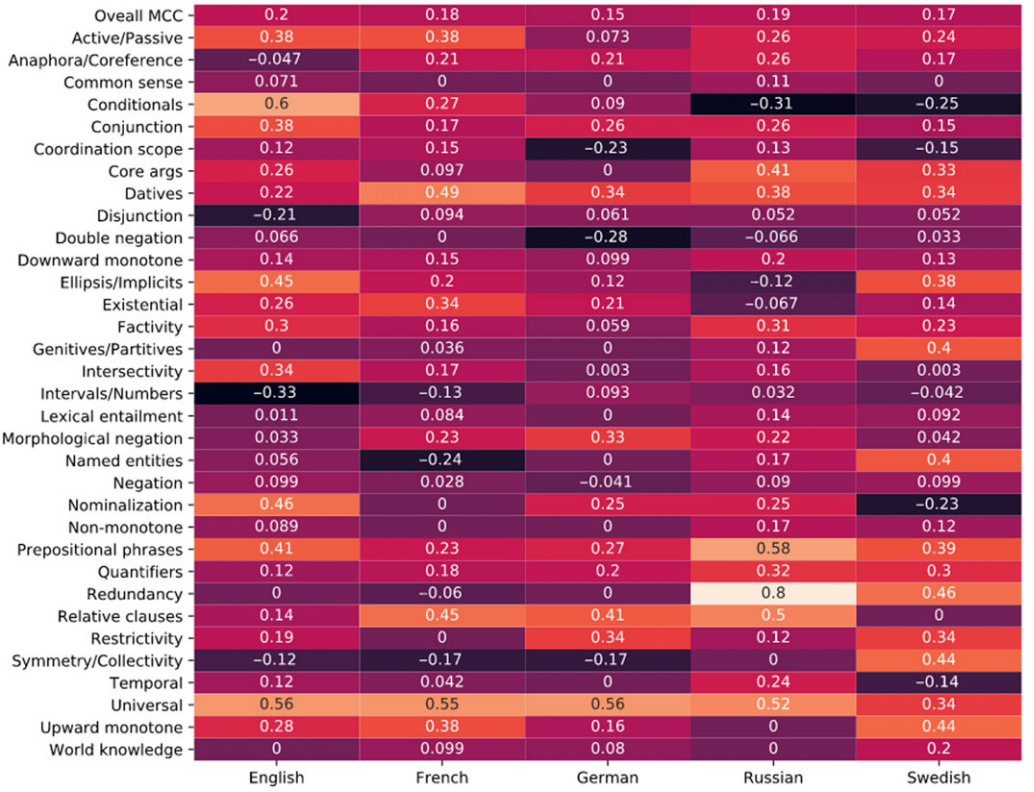

Figure 1. Heat map of the mBERT's language-wise evaluation on the diagnostic datasets. The brighter the color, the higher the MCC score.

resulting Pearson's correlation is 0.3, which denotes that the knowledge obtained during fine-tuning predominantly varies across the languages, and there is no general pattern in the model behavior. For instance, *Conditionals* contribute to the correct predictions for English (MCC = 0.6), slightly lower for French (MCC = 0.27), are neutral for German (MCC = 0.09) and do not help to solve the task for Russian (MCC = −0.31) and Swedish (MCC = −0.25). On the other hand, some features receive similar MCC scores for specific languages, such as *Active/Passive* (**English**: MCC = 0.38; **French**: MCC = 0.38; **Russian**: MCC = 0.26; **Swedish**: MCC = 0.24), *Anaphora/Coreference* (**French**: MCC = 0.21; **German**: MCC = 0.21; **Russian**: MCC = 0.26), *Common sense* (**French**: MCC = 0; **German**: MCC = 0; **Swedish**: MCC = 0), *Datives* (**German**: MCC = 0.34; **Russian**: MCC = 0.38; **Swedish**: MCC = 0.34), *Genitives/Partitives* (**English**: MCC = 0; **French**: MCC = 0.036; **German**: MCC = 0), and *Symmetry/Collectivity* (**English**: MCC = −0.12; **French**: MCC = −0.17; **German**: MCC = −0.17).

5.2 Fine-tuning stability and random seeds

We fine-tune the mBERT model multiple times while changing *only* the random seeds ∈ [0;5] for each considered language as described in Section 4. Figure 2 shows the seed-wise results for English. The results for the other languages are presented in Appendix 8.1. The overall pattern is that the correlation of the fine-grained diagnostic features and model outputs varies w.r.t. the random seed. Namely, some features demonstrate a large variance in the MCC score over different random seeds, for example, *Conditionals* (**English**: MCC = 0.6 [0]; MCC = 0.13 [1, 4, 5]),

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

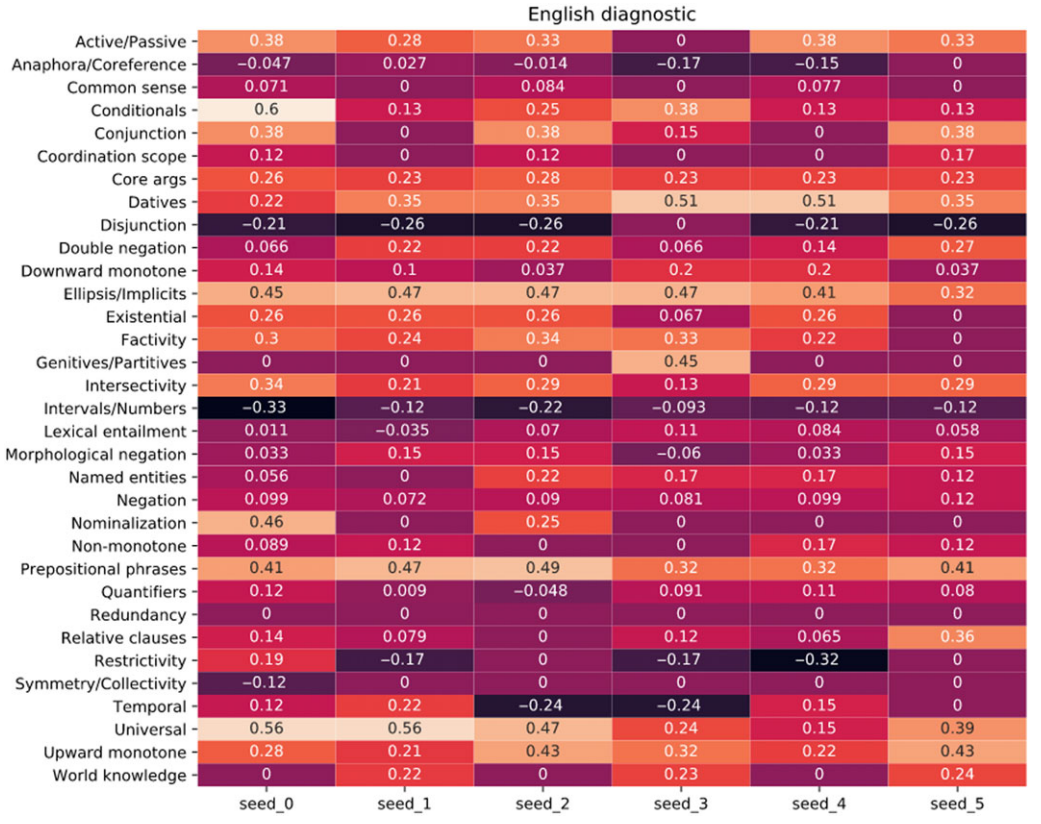

Figure 2. MCC scores on the English diagnostic dataset for mBERT fine-tuned with multiple random seeds.

Nominalization (**English**: MCC = 0.46 [0]; MCC = 0.46 [1, 3, 4, 5]), *Datives* (**French**: MCC = 0.64 [4]; MCC = 0.76 [5]; MCC = 0 [1, 3]), *Non-monotone* (**French**: MCC = 0 [0, 2]; MCC = −0.58 [4]; MCC = 0.21 [5]), *Genitives/Partitives* (**German**: MCC = 0 [0, 1]; MCC = 0.56 [2]; MCC = −0.29 [4]), *Restrictivity* (**Russian**: MCC = 0.12 [0, 2, 5]; MCC = 0 [3, 4]; MCC = −0.65 [1]), and *Redundancy* (**Swedish**: MCC = 0.34 [2]; MCC = 0 [3]; MCC = 0.8 [5]). On the one hand, a number of features positively correlates with the model predictions regardless the random seed, such as *Core args*, *Intersectivity*, *Prepositional phrases*, *Datives* (**English**); *Active/Passive*, *Existential*, *Upward monotone* (**French**); *Anaphora/Coreference* and *Universal* (**German**); *Factivity* and *Redundancy* (**Russian**); *Symmetry/Collectivity* and *Upward monotone* (**Swedish**). Some features, on the other hand, predominantly receive negative MCC scores: *Disjunction* and *Intervals/Numbers* (**English**), *Symmetry/Collectivity* (**French** and **Russian**), *Coordination scope* and *Double negation* (**German**), *Conditionals* and *Temporal* (**Swedish**). Table 3 aggregates the results of the seed-wise diagnostic evaluation for each language. While overall MCC scores within each language insignificantly differ, the mBERT model still have a weak correlation with the linguistic properties. Besides, the pairwise Pearson's correlation coefficients between the RS models[d] vary between languages up to 0.22, which specifies that fine-tuning stability of the mBERT model is dependent upon language.

Table 6. (see Appendix 8.1) presents granular results of the *per-category* fine-tuning stability of the mBERT model for each language. We now describe the categories that have received the

[d]We refer to the *RS model* as the model instance fine-tuned with a specific random seed value.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

Table 3. Results of the fine-tuning stability experiments w.r.t. random seeds for each language. **Overall MCC** = overall MCC scores of each RS model averaged by the total number of RS models. **RS corr.** = pairwise Pearson's correlation coefficients between the RS models' MCC scores, averaged by the total number of random seed pairs.

Language	Overall MCC	RS corr.
English	0.200 ± 0.016	0.634
French	0.178 ± 0.027	0.529
German	0.158 ± 0.024	0.411
Russian	0.182 ± 0.033	0.455
Swedish	0.169 ± 0.028	0.517
Average	0.177 ± 0.026	0.509

less and most significant standard deviations in the MCC scores over multiple random seeds. For most of the languages, the most stable categories are *Common sense* ($\sigma \in [0.04; 0.09]$) and *Factivity* ($\sigma \in [0.04; 0.1]$), while the most unstable ones are the categories of the *Lexical Semantics*, *Logic* and *Predicate-Argument Structure*, for example, *Genitives/Partitives* ($\sigma \in [0.17; 0.31]$), *Datives* ($\sigma \in [0.12; 0.34]$), *Restrictivity* ($\sigma \in [0.04; 0.3]$), and *Redundancy* ($\sigma \in [0.16; 0.32]$). The variance in the performance indicates the inconsistency of the linguistic generalization on a certain group of categories both collectively and discretely for the languages.

5.3 Fine-tuning stability and dataset size

Recent and contemporaneous studies report that a small number of training samples leads to unstable fine-tuning of the BERT model (Devlin et al., 2019; Phang et al., 2018; Zhu et al., 2019; Pruksachatkun et al., 2020a; Dodge et al., 2020). Toward that end, we conduct two experiments to investigate how additional training data impacts the fine-tuning stability in the cross-lingual transfer setting and how it changes while the number of training samples gradually increases. We use the MNLI (Williams et al., 2018) dataset for English and collapse *"neutral"* and *"contradiction"* samples into the *"not entailment"* label to meet the format of the RTE task (Wang et al., 2019). The resulting number of the additional training samples is 374k which are added to each language's corresponding RTE training data.

Does extra data in English improve stability for all languages? To analyze the performance patterns, we compute deltas between the feature-wise MCC scores and standard deviation values (σ) when using a single RTE training dataset (see Section 5.2) and a combination of the RTE and MNLI training datasets. Figure 3 shows heat maps of how the fine-tuning stability has changed after fine-tuning on the additional data. We find that the MCC scores have increased for 32% categories among all languages on average (delta between the MCC scores is more than 0.1). The per-category fine-tuning stability has improved for 34% of categories among all languages on average (delta between the σ values is below -0.05)[e]. An interesting observation is that some categories receive confident performance improvements for *all* languages (the MCC delta is above 0.2). Such categories include *Conjunction*, *Coordination scope*, *Genitives/Partitives*, *Non-monotone*, *Prepositional phrases*, *Redundancy*, and *Relative clauses*. However, the additional data

[e]The percentage corresponds to the fraction of the heat map cell values for all languages that are higher/lower than a specified threshold for the corresponding metric. The thresholds are chosen empirically and can be adjusted depending on the strictness of the experimental setting.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

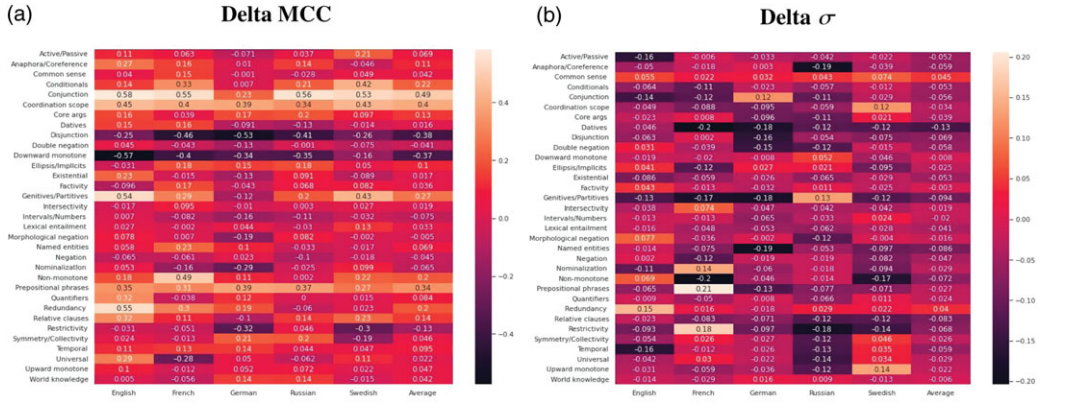

Figure 3. Feature-wise heat maps of the performance patterns after fine-tuning on combined RTE and MNLI training datasets. **Left**: Delta between MCC scores (higher is better). **Right**: Delta between standard deviation values (lower is better).

does not help for learning the *Disjunction* and *Downward monotone* categories and even hurts the performance as opposed to the results in Section 5.2. We also find that 61% of categories for Russian have the σ deltas below -0.05, indicating that the per-category stability can be greatly improved by extending the training data with examples in the English language.

Table 4 presents the results of this setting with a comparison to the previous experiments where the model is fine-tuned on the standardized train data size with multiple random seeds (see Section 5.1 and 5.2). The overall trend is that extension of the RTE training data with the MNLI samples helps to improve the fine-tuning stability for each language. Overall MCC scores for the diagnostic features have increased from 0.177 to 0.263 on average (up by 49%), and the average standard deviation decreased by 0.166. Analyzing the impact on the fine-tuning stability w.r.t. random seed (see Appendix 8.2), we observe that variance in the MCC scores between the RS models has predominantly decreased for all languages. Moreover, pairwise Pearson's correlation coefficients between the RS models have improved from 0.509 to 0.837 on average (up by 64%).

How many training samples are required for stability? To investigate the fine-tuning stability in the context of the training data size, we fine-tune the mBERT model as described in Section 4, while changing random seed $\in [0; 5]$ and gradually adding the MNLI samples $\in [1k, 5k, 10k, 50k, 100k, 200k, 250k, 374k]$ to the RTE training data for English and Russian. Figure 4 shows the results of this experiment. Despite the fact that the overall MCC scores stop increasing at the size of $RTE + 10k$ for both languages, the *RS corr.* is steadily improving, indicating a smaller variance in the MCC scores between the RS models. Besides, the model needs more data to improve the stability for Russian (recall that we add extra data in English).

5.4 Fine-tuning stability and presence of linguistic categories

We conduct the following experiment to investigate the relationship between the fine-tuning stability and particular diagnostic categories in the training data. We design a rule-based pipeline for annotating 15 out of 33 diagnostic features for English and Russian. Then, we evaluate the model depending on their presence percentage in the corresponding RTE training dataset combined with $10k$ training samples from MNLI (this amount of extra data is selected based on the results in Section 5.3.).

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

Table 4. Results of the fine-tuning stability w.r.t using additional MNLI training samples in the cross-lingual transfer setting. **Overall MCC** = overall MCC scores of each RS model averaged by the total number of RS models. **RS corr.** = pairwise Pearson's correlation coefficients between the RS models' MCC scores, averaged by the total number of random seed pairs.

Language	Fine-tuning data	Overall MCC	RS corr.
English	**RTE**	0.200 ± 0.016	0.634
	RTE & MNLI	0.294 ± 0.006	0.929
French	**RTE**	0.178 ± 0.027	0.529
	RTE & MNLI	0.268 ± 0.010	0.822
German	**RTE**	0.158 ± 0.024	0.411
	RTE & MNLI	0.213 ± 0.010	0.836
Russian	**RTE**	0.182 ± 0.033	0.455
	RTE & MNLI	0.263 ± 0.012	0.810
Swedish	**RTE**	0.169 ± 0.028	0.517
	RTE & MNLI	0.277 ± 0.016	0.785
Average	**RTE**	0.177 ± 0.177	0.509
	RTE & MNLI	0.263 ± 0.011	0.836

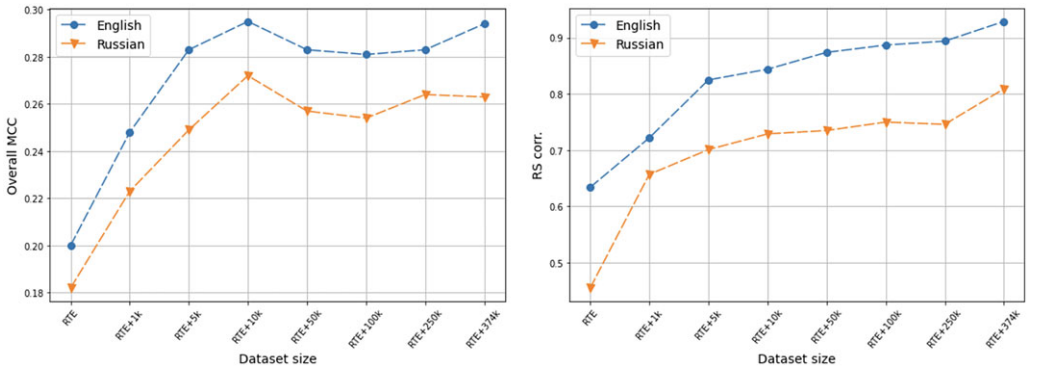

Figure 4. Results of the fine-tuning stability w.r.t. the number of additional MNLI training samples added to the RTE training data for English and Russian. **Overall MCC** = overall MCC scores of each RS model averaged by the total number of RS models. **RS corr.** = pairwise Pearson's correlation coefficients between the RS models' MCC scores, averaged by the total number of random seed pairs.

Description of annotation pipeline. Our study suggests that annotation of low-level diagnostic categories can be partially automatized based on features expressed lexically or grammatically. *Lexical Semantics* can be detected by the presence of quantifiers, negation morphemes, factivity verbs, and proper nouns. *Logic* features can be expressed with the indicators of temporal relations (mostly prepositions, conjunctions, particles, and deictic words), negation, and conditionals. Features from the *Predicate-Argument Structure* category can be identified with pronouns and syntactic tags (e.g., *Relative clauses*, *Datives*, etc.). However, *Knowledge* categories cannot be obtained in this manner.

Such approach relies only on the surface representation of the feature and is limited by the coverage of the predefined rules, thus giving space to false-negative results. Keeping this in mind,

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

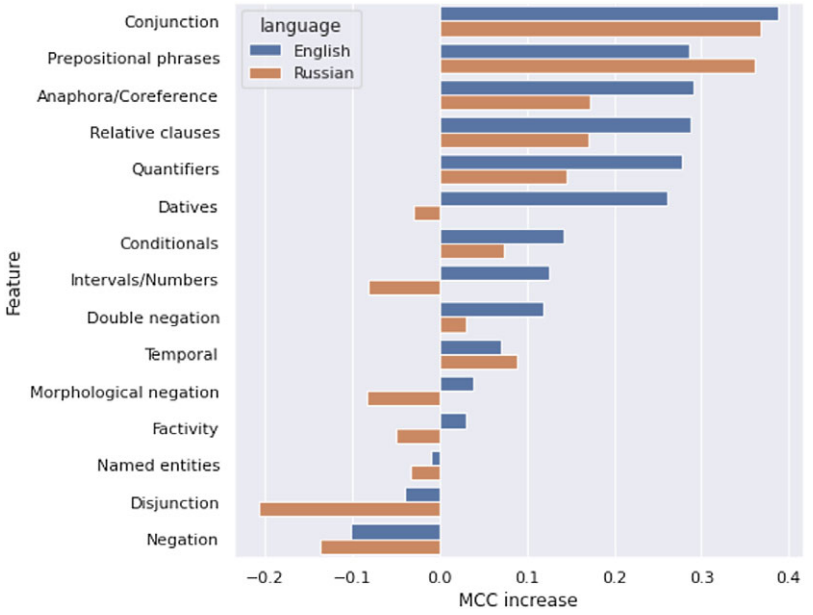

Figure 5. Distribution of the model MCC scores when fine-tuned on the combined data (**RTE** + 10k) as opposed to the standardized dataset size.

we construct a set of linguistic heuristics to identify the presence of a particular feature based on the morphosyntactic and NER annotation with spaCy[f] for English, and built-in dictionaries and morphological analysis with pymorphy2 for Russian (Korobov 2015). We also construct specific word lists for most of the features for both languages, for example, *"all," "some," "every," "any," "anyone," "everyone," "nothing*," etc. (*Quantifiers*). The heuristics for the Russian language have several differences. For instance, dative constructions are detected by the morphological analysis of the nouns or pronouns, as the case is explicitly expressed in the flexion.

Stability and category distribution. We use the pipeline to annotate each training sample from RTE, TERRa, and the MNLI 10k subset. Table 7 presents the feature distributions for the datasets (see Appendix 8.3). Figure 5 depicts the model performance trajectories when fine-tuned on the combined data as opposed to the standardized dataset size (see Section 5.1). The behavior is predominantly similar for both languages, and there is a strong correlation of 0.94 between the MCC performance improvements. We select four features for further analysis[g]: *Conjunction* (the MCC score improved for both languages), *Anaphora/Coreference* (there is a significant difference in the feature distribution between RTE and MNLI, and no such difference between TERRa and MNLI), *Negation* (the MCC score decreased for both languages, and the feature distribution differs between the languages), and *Disjunction* (the MCC score decreased for both languages). For each considered feature, we construct three *controllable* subsets with a varying percentage of the presence in the training data. We follow the same fine-tuning and evaluation strategy (see Section 4), changing random seed $\in [0;5]$ and the feature percentage presence $\in [25, 50, 75]$. Table 5 presents the results of the experiment. The general pattern observed for both languages is that adding more feature-specific training samples may rather hurt the fine-tuning stability along with the MCC score for the feature.

[f] https://spacy.io/
[g] Our future work includes analysis of the other features, specifically for French, German, and Swedish.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

Table 5. Results of the fine-tuning stability w.r.t. varying degree of the feature distribution in the MNLI subset for English and Russian. **Feature MCC** = feature MCC score of each RS model averaged by the total number of RS models. **RS corr.** = pairwise Pearson's correlation coefficients between the RS models' MCC scores, averaged by the total number of random seed pairs.

		Feature MCC		RS corr.	
Feature	**Presence, %**	**En**	**Ru**	**En**	**Ru**
Conjunction	25	0.717 ± 0.07	0.656 ± 0.05	0.812	0.732
	50	0.752 ± 0.03	0.648 ± 0.12	0.783	0.749
	75	0.682 ± 0.01	0.534 ± 0.13	0.792	0.684
Negation	25	0.013 ± 0.06	-0.032 ± 0.04	0.812	0.712
	50	0.014 ± 0.05	0.005 ± 0.06	0.839	0.751
	75	0.004 ± 0.05	0.029 ± 0.08	0.742	0.684
Anaphora/coreference	25	0.125 ± 0.05	0.125 ± 0.05	0.845	0.845
	50	0.171 ± 0.05	0.223 ± 0.08	0.848	0.666
	75	0.202 ± 0.05	0.197 ± 0.04	0.778	0.704
Disjunction	25	-0.327 ± 0.16	-0.078 ± 0.14	0.841	0.706
	50	-0.198 ± 0.05	-0.175 ± 0.18	0.781	0.752
	75	-0.146 ± 0.06	-0.092 ± 0.8	0.831	0.608

Feature MCC. The highest MCC scores for English are achieved when adding 50% (*Conjunction*, *Negation*), or 75% extra samples (*Anaphora/Coreference*, *Disjunction*). In contrast, this amount of data has decreased the MCC performance for Russian (*Conjunction*, *Negation*). Instead, the minimum number of 25% additional samples are required to receive the best MCC scores for the categories of *Conjunction* and *Disjunction*. *Negation* obtains an insignificant improvement when adding 75% samples, and *Anaphora/Coreference* is of 0.223 MCC at 50% extra data.

Fine-tuning stability. Despite the fact that the feature MCC scores may increase, the fine-tuning stability may decrease for the identical amounts of additional training samples, for example, *Conjunction* (English and Russian), *Negation* (Russian), *Anaphora/Coreference* (English), and *Disjunction* (English and Russian). The minor variance between the RS models is predominantly the 25% or 50% extra data size for both languages.

Probing analysis. To analyze from another perspective, we apply the annotation pipeline to construct three probing tasks, aimed at identifying the presence of categories of *Logic*, *Lexical Semantics*, and *Predicate-Argument structure*. More details can be found in Appendix 8.4.

6. Discussion

Acquiring linguistic knowledge through NLI. A thorough language-wise analysis using the proposed multilingual datasets reveals how well the model learns the phenomena it is intended to learn for solving the NLI task. Despite the variability in the MCC performance, mBERT shows

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

a similar behavior on a number of features on the languages that differ in their richness of morphology and syntax (see Section 5.1). Specifically, the model outputs are positively correlated with the following diagnostic categories that reflect the language peculiarities: *Logic* (*Upward monotone*, *Conditionals*, *Existential*, *Universal*, and *Conjunction*), *Lexical semantics* (*Named entities*), and *Predicate-Argument structure* (*Ellipsis*, *Coordination scope*, and *Anaphora/Coreference*). On the contrary, there is a number of features that predominantly receive negative MCC scores: *Logic* (*Disjunction*, *Downward monotone*, and *Intervals/Numbers*) and *Predicate-Argument structure* (*Restrictivity*). The *Logic* features are reminiscent of the properties of formal semantics, which captures the meaning of linguistic expressions through their logical interpretation utilizing formal models (Venhuizen et al., 2021). Monotonicity (*Upward/Downward monotone*), as one of such features, covers various systematic patterns and allows for assessing inferential systematicity in natural languages. In line with (Yanaka et al., 2019b), our results show that the model generally struggles to learn the *Downward* monotone inferences with *Disjunction* for all languages. Another phenomenon to which mBERT is insensitive is the category of *Negation*. The model outputs weakly correlate with the true labels when the sample contains *Negation*, *Double negation*, and *Morphological negation*, indicating that the model fails to infer this core construction, which is a well-studied problem in the field (Naik et al., 2018; Ettinger 2020; Hosseini et al., 2021). Recently, Wallace et al., (2019) have shown that it is difficult for contextualized LMs to generalize beyond the numerical values seen during training, and various datasets and model improvements have been proposed to analyze and enhance the understanding of numeracy (Thawani et al., 2021). The results for the category *Intervals/Numbers* in the context of the NLI problem reveal that numerical reasoning does not correlate with the expected model behavior (German and Russian) and even confuses the model (English, French, and Swedish). We also find that the results for the category of *Symmetry/Collectivity* (*Lexical Semantics*) vary between the considered languages, achieving negative MCC scores for most of them (English, French, and German). We relate this to the fact that the model may overly rely on the knowledge about entities and relations between them, refined from the pretraining corpora, so that linguistic expressions of the features are ignored (Tanchip et al., 2020; Kassner and Schütze 2020). Last but not least, we find that broadly defined categories such as *Common sense* and *World knowledge* do not show a significant correlation for all analyzed languages.

Comparing our results with the diagnostic evaluation of Chinese Transformer-based models on the NLI task (Xu et al., 2020), we observe the following similar trends[h]. Consistent with our findings, *Common sense* and *Monotonicity* appear to be quite challenging to learn. However, the results for low-level categories that fall under *Predicate-Argument Structure* might differ. While the Chinese LMs achieve an average accuracy score of 58% on this category, mBERT has a hard time dealing with *Nominalization* or *Restrictivity* but tends to learn *Coordination scope*, *Prepositional phrases*, and *Genitives/Partitives*. At the same time, predictions of mBERT weakly correlate with *Double negation*, but the Chinese models receive an average accuracy score of 60%. Similarly, *Lexical semantics* is one of the best-learned Chinese categories; however, the mBERT model does not demonstrate a consistent behavior on the corresponding low-level categories. A more detailed investigation of cross-lingual LMs on these typologically diverse languages may shed light on how the models learn linguistic properties crucial for the NLI task and provide more insights on the cross-lingual transfer of language-specific categories and markers (Hu et al., 2021).

The impact of random seeds. Our results are consistent with McCoy et al., (2020) who find that the instances of BERT fine-tuned on MNLI vary widely in their performance on the HANS dataset. In our work, the examination of the mBERT's performance on the diagnostic datasets reveals a significant variance in the MCC scores and standard deviation w.r.t. random seeds for the majority of considered languages (see Section 5.2, Appendix 8.1). We observe significant standard deviations

[h]Note that the results are not directly comparable in terms of target metrics, dataset domains, and models.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

in the diagnostic performance, which indicates both *per-language* and *per-category* fine-tuning instability of the mBERT model. The findings highlight the importance of evaluating models on multiple restarts, as the scores obtained by a single model instance may not extrapolate to other instances, specifically in the multilingual benchmarks such as XGLUE (Liang et al., 2020) and XTREME (Hu et al., 2020b). Namely, the features that are crucial for diagnostic analysis of LMs might not be appropriately learned by a particular instance, which may underscore their generalization abilities on the canonical leaderboards or even question whether LMs are indeed capable of capturing them either from pretraining or fine-tuning data. The statements are supported by the probing analysis, which shows that fine-tuning of mBERT on the RTE tasks with varying random seeds may unpredictably affect the model's knowledge (see Appendix 8.4). Specifically, the effect can be abstracted as twofold: fine-tuned mBERT model either "forget" about a peculiar linguistic category, or "acquire" the uncertain knowledge which is demonstrated by sharp increases and decreases in the probe performance over several languages (Singh et al., 2020).

The impact of dataset size and feature proportions. Prior studies have reported contradictory results about the effect of adding/augmenting training data on the linguistic generalization and inference capabilities of LMs. Some works demonstrate that counterfactually augmented data does not yield generalization improvements on the NLI task (Huang et al., 2020). However, most recent studies show that fine-tuning BERT on additional NLI samples that cover particular inference features improves their understanding while retaining or increasing the downstream performance on NLI benchmarks (Yanaka et al., 2020, 2019b; Richardson et al., 2020; Min et al., 2020; Hosseini et al., 2021). Besides, the proportion of the features in the training data can be crucial for the model performance (Yanaka et al., 2019a). One of the closely related works by (Hu et al., 2021) tests cross-lingual transfer abilities of XLM-R (Conneau et al., 2020) on the NLI task for Chinese, exploring configurations of fine-tuning the model on combinations of Chinese and English data and evaluating it on diagnostic datasets. Particularly, the model achieves the best performance when fine-tuned on concatenated OCNLI (Hu et al., 2020a) and English NLI datasets (e.g., Bowman et al., 2015; Williams et al., 2018; Nie et al., 2020) on the majority of covered diagnostic features, including uniquely Chinese ones: *Idioms*, *Non-core argument*, *Pro-drop*, *Time of event*, *Anaphora*, *Argument structure*, *Comparatives*, *Double negation*, *Lexical semantics*, and *Negation*. The results suggest that XLM-R can learn meaningful linguistic representations beyond surface properties and even strengthen the knowledge with the transfer from English, outperforming its monolingual counterparts.

Consistent with the latter studies, we find that extra data only in English provides better generalization capabilities of mBERT for all considered languages, which differ in their peculiarities of morphology and syntax. We also observe that using additional English data improves the fine-tuning stability, resulting in lower standard deviation values and higher Pearson's correlation between the model instances' scores (see Section 5.3). Another finding is that the number of training examples containing a particular feature might be critical for both diagnostic performance and fine-tuning stability of the mBERT model (see Section 5.4).

Limitations. The concept of benchmarking has become a standard paradigm for evaluating LMs against one another and human solvers, and dataset design protocols for the other languages are generally reproduced from English. However, there are still several methodological concerns, one of which is the dataset design and annotation choices (Rogers 2019; Dehghani et al., 2021). It should be noted that a relatively small number of dataset samples has a common basis in benchmarking due to expensive annotation or the need for expert competencies. Unlike datasets for machine-reading comprehension, such as MultiRC (Khashabi et al., 2018) and ReCoRD (Zhang et al., 2018), the GLUE-style datasets for learning choice of alternatives, logic, and causal relationships are often represented by a smaller number of manually collected and verified samples. They are by design sufficient for the human type of generalization but often pose a challenge

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

for the tested LMs. The broad-coverage diagnostic dataset is standard practice for assessing linguistic generalization of LMs. Nevertheless, it contains 1104 samples, and the number of samples for certain features includes only 14 samples (*Universal* and *Existential*). These dataset design choices might not provide an opportunity for a fair comparison and reliable interpretation of LMs, which might be supported by bootstrap techniques or construction of evaluation sets balanced by the number of analyzed phenomena. Evaluating datasets for sufficiency for in-distribution and out-of-distribution generalization is another relevant challenge in the field. The solution might significantly help both in interpreting model learning outcomes and in designing better evaluation suites and benchmarks. Recall that our results might not be transferable to other multilingual models, specifically different in the architecture design and pretraining objectives, for example, XLM-R, mBART (Liu et al., 2020), and mT5 (Xue et al., 2021).

7. Conclusion

This paper presents an extension of the ongoing research on the fine-tuning stability and consistency of linguistic generalization to the multilingual setting. We propose six GLUE-style textual entailment and broad-coverage diagnostic datasets for French, German, and Swedish. The datasets are constructed by translating the original datasets for English and Russian, with culture-specific phenomena localized and language phenomena adapted under linguistic expertise. We address the problem in the NLI task and analyze the linguistic competence of the mBERT model along with the impact of the random seed choice, training data size, and presence of linguistic categories in the training data. The method includes the standard SuperGLUE fine-tuning and evaluation procedure, and we ensure that the model is run with precisely the same hyperparameters but with different random seeds. The mBERT model demonstrates the *per-category* instability generally for categories that involve lexical semantics, logic, and predicate-argument structure and struggles to learn monotonicity, negation, numeracy, and symmetry. However, related languages show similar performance in active and passive voice, conjunction, disjunction, prepositional phrases, and quantifiers. We also find that the generalization performance and fine-tuning stability can be improved for all languages by using additional data only in English, contributing to the cross-lingual transfer capabilities of multilingual LMs. However, the number of training samples containing a particular feature might also hurt all model instances' performance. We leave a more detailed investigation of this behavior for future work. Another fruitful direction is analyzing a more diverse set of monolingual and multilingual LMs, varying by the architecture design and pretraining objectives. In general, our results are consistent with a growing body of related studies which explore aspects of learning inference properties from different perspectives, including findings for Chinese, a language typologically different from the considered ones in our work. We are publicly releasing the datasets, hoping to foster the diagnostic investigation of LMs in a cross-lingual scenario, particularly in terms of benchmarking, which might promote a more holistic understanding of multilingualism in LMs and their cross-lingual knowledge transfer abilities.

Funding statement. The work has been supported by the Ministry of Science and Higher Education of the Russian Federation within Agreement No 075-15-2020-793.

References

Al-Shabab O. (1996). Interpretation and the language of translation: creativity and conventions in translation.

Belinkov Y., **Poliak A.**, **Shieber S.**, **Van Durme B.** and **Rush A.** (2019). Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 877–891.

Bengio Y. (2012). Practical recommendations for gradient-based training of deep architectures.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

Bentivogli L., **Clark P.**, **Dagan I.** and **Giampiccolo D.** (2009). The fifth pascal recognizing textual entailment challenge. In *TAC*.

Bhojanapalli S., **Wilber K.**, **Veit A.**, **Rawat A.S.**, **Kim S.**, **Menon A.** and **Kumar S.** (2021). On the reproducibility of neural network predictions. arXiv preprint arXiv:2102.03349.

Bowman S.R., **Angeli G.**, **Potts C.** and **Manning C.D.** (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics, pp. 632–642.

Conneau A., **Khandelwal K.**, **Goyal N.**, **Chaudhary V.**, **Wenzek G.**, **GuzmÁn F.**, **Grave E.**, **Ott M.**, **Zettlemoyer L.** and **Stoyanov V.** (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8440–8451.

Conneau A., **Kruszewski G.**, **Lample G.**, **Barrault L.** and **Baroni M.** (2018a). What you can cram into a single $&!#∗ vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume* 1: *Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 2126–2136.

Conneau A., **Rinott R.**, **Lample G.**, **Williams A.**, **Bowman S.**, **Schwenk H.** and **Stoyanov V.** (2018b). XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 2475–2485.

Cui L., **Cheng S.**, **Wu Y.** and **Zhang Y.** (2020). Does bert solve commonsense task via commonsense knowledge?

Dagan I., **Glickman O.** and **Magnini B.** (2005). The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*. Springer, pp. 177–190.

Dehghani M., **Tay Y.**, **Gritsenko A.A.**, **Zhao Z.**, **Houlsby N.**, **Diaz F.**, **Metzler D.** and **Vinyals O.** (2021). The benchmark lottery.

Devlin J., **Chang M.-W.**, **Lee K.** and **Toutanova K.** (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)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186.

Dodge J., **Ilharco G.**, **Schwartz R.**, **Farhadi A.**, **Hajishirzi H.** and **Smith N.** (2020). Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. arXiv preprint arXiv:2002.06305.

Ettinger A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, **8**, 34–48.

Giampiccolo D., **Magnini B.**, **Dagan I.** and **Dolan B.** (2007). The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*. Prague: Association for Computational Linguistics, pp. 1–9.

Glockner M., **Shwartz V.** and **Goldberg Y.** (2018). Breaking NLI systems with sentences that require simple lexical inferences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume* 2: *Short Papers)*. Melbourne, Australia: Association for Computational Linguistics, pp. 650–655.

Goldberg Y. (2019). Assessing BERT's syntactic abilities.

Gorodkin J. (2004). Comparing two k-category assignments by a k-category correlation coefficient. *Computational Biology and Chemistry* **28**(5–6), 367–374.

Haim R.B., **Dagan I.**, **Dolan B.**, **Ferro L.**, **Giampiccolo D.**, **Magnini B.** and **Szpektor I.** (2006). The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

He P., **Liu X.**, **Gao J.** and **Chen W.** (2021). Deberta: Decoding-enhanced bert with disentangled attention.

Henderson P., **Islam R.**, **Bachman P.**, **Pineau J.**, **Precup D.** and **Meger D.** (2018). Deep reinforcement learning that matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32.

Hossain, **M.M.**, **Kovatchev V.**, **Dutta P.**, **Kao T.**, **Wei E.** and **Blanco E.** (2020). An analysis of natural language inference benchmarks through the lens of negation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 9106–9118.

Hosseini A., **Reddy S.**, **Bahdanau D.**, **Hjelm R.D.**, **Sordoni A.** and **Courville A.** (2021). Understanding by understanding not: Modeling negation in language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 1301–1312.

Hu, **H.**, **Richardson**, **K.**, **Xu**, **L.**, **Li**, **L.**, **Kübler**, **S.** and **Moss**, **L.S.** (2020a). Ocnli: Original chinese natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP* 2020, pp. 3512–3526.

Hu H., **Zhou H.**, **Tian Z.**, **Zhang Y.**, **Patterson Y.**, **Li Y.**, **Nie Y.** and **Richardson K.** (2021). Investigating transfer learning in multilingual pre-trained language models through Chinese natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP* 2021. Online: Association for Computational Linguistics, pp. 3770–3785.

Hu J., **Ruder S.**, **Siddhant A.**, **Neubig G.**, **Firat O.** and **Johnson M.** (2020b). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

Hua H., **Li X.**, **Dou D.**, **Xu C.** and **Luo J.** (2021). Noise stability regularization for improving BERT fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 3229–3241.

Huang W., **Liu H.** and **Bowman S.R.** (2020). Counterfactually-augmented SNLI training data does not yield better generalization than unaugmented data. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*. Online: Association for Computational Linguistics, pp. 82–87.

Jawahar G., **Sagot B.** and **Seddah D.** (2019). What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3651–3657.

Kassner N. and **Schütze H.** (2020). Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7811–7818.

Khashabi D., **Chaturvedi S.**, **Roth M.**, **Upadhyay S.** and **Roth D.** (2018). Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, *Volume* 1 *(Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 252–262.

Kingma D.P. and **Ba J.** (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Korobov M. (2015). Morphological analyzer and generator for Russian and Ukrainian languages. In *International Conference on Analysis of Images, Social Networks and Texts AIST 2015: Analysis of Images, Social Networks and Texts*, vol. 542, pp. 320–332.

Kovaleva O., **Romanov A.**, **Rogers A.** and **Rumshisky A.** (2019). Revealing the dark secrets 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 4365–4374.

Le H., **Vial L.**, **Frej J.**, **Segonne V.**, **Coavoux M.**, **Lecouteux B.**, **Allauzen A.**, **Crabbé B.**, **Besacier L.** and **Schwab D.** (2020). FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 2479–2490.

Lee C., **Cho K.** and **Kang W.** (2019). Mixout: Effective regularization to finetune large-scale pretrained language models. arXiv preprint arXiv:1909.11299.

Liang Y., **Duan N.**, **Gong Y.**, **Wu N.**, **Guo F.**, **Qi W.**, **Gong M.**, **Shou L.**, **Jiang D.**, **Cao G.**, **Fan X.**, **Zhang R.**, **Agrawal R.**, **Cui E.**, **Wei S.**, **Bharti T.**, **Qiao Y.**, **Chen J.-H.**, **Wu W.**, **Liu S.**, **Yang F.**, **Campos D.**, **Majumder R.** and **Zhou M.** (2020). XGLUE: A new benchmark datasetfor cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 6008–6018.

Liška A., **Kruszewski G.** and **Baroni M.** (2018). Memorize or generalize? searching for a compositional rnn in a haystack. arXiv preprint arXiv:1802.06467.

Liu Y., **Gu J.**, **Goyal N.**, **Li X.**, **Edunov S.**, **Ghazvininejad M.**, **Lewis M.** and **Zettlemoyer L.** (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics* **8**, 726–742.

Loshchilov I. and **Hutter F.** (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Madhyastha P. and **Jain R.** (2019). On model stability as a function of random seed. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Hong Kong, China: Association for Computational Linguistics, pp. 929–939.

Marelli M., **Menini S.**, **Baroni M.**, **Bentivogli L.**, **Bernardi R.** and **Zamparelli R.** (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Reykjavik, Iceland: European Language Resources Association (ELRA), pp. 216–223.

McCoy R.T., **Frank R**, and **Linzen T.** (2018). Revisiting the poverty of the stimulus: hierarchical generalization without a hierarchical bias in recurrent neural networks.

McCoy R.T., **Min J.** and **Linzen T.** (2020). BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Online: Association for Computational Linguistics, pp. 217–227.

McCoy T., **Pavlick E.** and **Linzen T.** (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3428–3448.

Merchant A., **Rahimtoroghi E.**, **Pavlick E.** and **Tenney I.** (2020). What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Online: Association for Computational Linguistics, pp. 33–44.

Miaschi A., **Brunato D.**, **Dell'Orletta F.** and **Venturi G.** (2020). Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 745–756.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

Min J., **McCoy R.T.**, **Das D.**, **Pitler E.** and **Linzen**, **T.** (2020). Syntactic data augmentation increases robustness to inference heuristics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 2339–2352.

Mosbach M., **Andriushchenko M.** and **Klakow D.** (2020a). On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. In *International Conference on Learning Representations*.

Mosbach M., **Khokhlova A.**, **Hedderich M.A.** and **Klakow D.** (2020b). On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Online: Association for Computational Linguistics, pp. 68–82.

Naik A., **Ravichander A.**, **Sadeh N.**, **Rose C.** and **Neubig G.** (2018). Stress test evaluation for natural language inference. In *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 2340–2353.

Nie Y., **Williams A.**, **Dinan E.**, **Bansal M.**, **Weston J.** and **Kiela D.** (2020). Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4885–4901.

Nisioi S., **Rabinovich E.**, **Dinu L.P.** and **Wintner S.** (2016). A corpus of native, non-native and translated texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 4197–4201.

Phang J., **Févry T.** and **Bowman S.R.** (2018). Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. arXiv preprint arXiv:1811.01088.

Pruksachatkun Y., **Phang J.**, **Liu H.**, **Htut P.M.**, **Zhang X.**, **Pang R.Y.**, **Vania C.**, **Kann K.** and **Bowman S.R.** (2020a). Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 5231–5247.

Pruksachatkun Y., **Yeres P.**, **Liu H.**, **Phang J.**, **Htut P. M.**, **Wang A.**, **Tenney I.** and **Bowman S.R.** (2020b). jiant: A software toolkit for research on general-purpose text understanding models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, pp. 109–117.

Raffel C., **Shazeer N.**, **Roberts A.**, **Lee K.**, **Narang S.**, **Matena M.**, **Zhou Y.**, **Li W.** and **Liu P.J.** (2020). Exploring the limits of transfer learning with a unified text-to-text transformer.

Richardson K., **Hu H.**, **Moss L.** and **Sabharwal A.** (2020). Probing natural language inference models through semantic fragments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8713–8721.

Rogers A. (2019). How the transformers broke nlp leaderboards.

Rogers A. (2021). Changing the world by changing the data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 2182–2194.

Rogers A., **Kovaleva O.** and **Rumshisky A.** (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics* **8**, 842–866.

Rybak P., **Mroczkowski R.**, **Tracz J.** and **Gawlik I.** (2020). KLEJ: Comprehensive benchmark for Polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1191–1201.

Sanchez I., **Mitchell J.** and **Riedel S.** (2018). Behavior analysis of NLI models: Uncovering the influence of three factors on robustness. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume* 1 *(Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1975–1985.

Shavrina T., **Fenogenova A.**, **Anton E.**, **Shevelev D.**, **Artemova E.**, **Malykh V.**, **Mikhailov V.**, **Tikhonova M.**, **Chertok A.** and **Evlampiev A.** (2020). RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 4717–4726.

Shavrina T. and **Shapovalova O.** (2017). To the methodology of corpus construction for machine learning:"taiga". syntax tree corpus and parser. *Corpus Linguistics 2017*, p. 78.

Singh J., **Wallat J.** and **Anand A.** (2020). BERTnesia: Investigating the capture and forgetting of knowledge in BERT. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. Online: Association for Computational Linguistics, pp. 174–183.

Storks S., **Gao Q.** and **Chai J.Y.** (2019). Recent advances in natural language inference: A survey of benchmarks, resources, and approaches. arXiv preprint arXiv:1904.01172.

Tanchip C., **Yu L.**, **Xu A.** and **Xu Y.** (2020). Inferring symmetry in natural language. In Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, pp. 2877–2886.

Thawani A., **Pujara J.**, **Ilievski F.** and **Szekely P.** (2021). Representing numbers in NLP: a survey and a vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 644–656.

Tsuchiya M. (2018). Performance impact caused by hidden bias of training data for recognizing textual entailment. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* Miyazaki, Japan: European Language Resources Association (ELRA).

Vaswani A., **Shazeer N.**, **Parmar N.**, **Uszkoreit J.**, **Jones L.**, **Gomez A.N.**, **Kaiser Ł.** and **Polosukhin I.** (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998–6008).

Venhuizen N.J., **Hendriks P.**, **Crocker M.W.** and **Brouwer H.** (2021). Distributional formal semantics. *Information and Computation*, p. 104763.

Wallace E., **Wang Y.**, **Li S.**, **Singh S.** and **Gardner M.** (2019). Do NLP models know numbers? probing numeracy in embeddings. 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 5307–5315.

Wang A., **Pruksachatkun Y.**, **Nangia N.**, **Singh A.**, **Michael J.**, **Hill F.**, **Levy O.** and **Bowman S.** (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In Advances in Neural Information Processing Systems, pp. 3266–3280.

Wang A., **Singh A.**, **Michael J.**, **Hill F.**, **Levy O.** and **Bowman S.** (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, pp. 353–355.

Warstadt A. and **Bowman S.R.** (2019). Linguistic analysis of pretrained sentence encoders with acceptability judgments. arXiv preprint arXiv:1901.03438.

Weber N., **Shekhar L.** and **Balasubramanian N.** (2018). The fine line between linguistic generalization and failure in Seq2Seq-attention models. In *Proceedings of the Workshop on Generalization in the Age of Deep Learning*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 24–27.

Williams A., **Nangia N.** and **Bowman S.** (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 1112–1122.

Wu J.M., **Belinkov Y.**, **Sajjad H.**, **Durrani N.**, **Dalvi F.** and **Glass J.** (2020). Similarity analysis of contextual word representation models. arXiv preprint arXiv:2005.01172.

Xu L., **Hu H.**, **Zhang X.**, **Li L.**, **Cao C.**, **Li Y.**, **Xu Y.**, **Sun K.**, **Yu D.**, **Yu C.**, **Tian Y.**, **Dong Q.**, **Liu W.**, **Shi B.**, **Cui Y.**, **Li J.**, **Zeng J.**, **Wang R.**, **Xie W.**, **Li Y.**, **Patterson Y.**, **Tian Z.**, **Zhang Y.**, **Zhou H.**, **Liu S.**, **Zhao Z.**, **Zhao Q.**, **Yue C.**, **Zhang X.**, **Yang Z.**, **Richardson K.** and **Lan Z.** (2020). CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 4762–4772.

Xue L., **Constant N.**, **Roberts A.**, **Kale M.**, **Al-Rfou R.**, **Siddhant A.**, **Barua A.** and **Raffel C.** (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 483–498.

Yanaka H., **Mineshima K.**, **Bekki D.** and **Inui K.** (2020). Do neural models learn systematicity of monotonicity inference in natural language? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 6105–6117.

Yanaka H., **Mineshima K.**, **Bekki D.**, **Inui K.**, **Sekine S.**, **Abzianidze L.** and **Bos J.** (2019a). Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Florence, Italy: Association for Computational Linguistics, pp. 31–40.

Yanaka H., **Mineshima K.**, **Bekki D.**, **Inui K.**, **Sekine S.**, **Abzianidze L.** and **Bos J.** (2019b). HELP: A dataset for identifying shortcomings of neural models in monotonicity reasoning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (∗SEM 2019)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 250–255.

Zhang S., **Liu X.**, **Liu J.**, **Gao J.**, **Duh K.** and **Durme B.V.** (2018). Record: Bridging the gap between human and machine commonsense reading comprehension.

Zhang Y., **Warstadt A.**, **Li X.**, and **Bowman S.R.** (2021). When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume* 1: *Long Papers)*. Online: Association for Computational Linguistics, pp. 1112–1125.

Zhao Y. and **Bethard S.** (2020). How does BERT's attention change when you fine-tune? an analysis methodology and a case study in negation scope. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4729–4747.

Zhu C., **Cheng Y.**, **Gan Z.**, **Sun S.**, **Goldstein T.** and **Liu J.** (2019). Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.

Zhuang D., **Zhang X.**, **Song S.L.** and **Hooker S.** (2021). Randomness in neural network training: Characterizing the impact of tooling. arXiv preprint arXiv:2106.11872.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

8. Appendix

8.1 Fine-tuning stability and random seeds

Table 6 presents the results of the *per-category* fine-tuning stability for each language.

Table 6. Results of the *per-category* fine-tuning stability for each language. The MCC scores are averaged over the total number of RS models. **Average** = The results averaged over five languages.

Feature	English	French	German	Russian	Swedish	Average
Active/passive	0.274 ± 0.16	0.245 ± 0.10	0.124 ± 0.13	0.310 ± 0.10	0.113 ± 0.09	0.213 ± 0.12
Anaphora/coreference	−0.071 ± 0.09	0.048 ± 0.13	0.260 ± 0.03	0.035 ± 0.23	0.228 ± 0.14	0.100 ± 0.12
Common sense	0.046 ± 0.04	0.076 ± 0.09	0.030 ± 0.04	0.066 ± 0.05	0.045 ± 0.08	0.053 ± 0.06
Conditionals	0.298 ± 020	0.040 ± 0.30	0.044 ± 0.09	−0.207 ± 0.14	−0.151 ± 0.18	0.004 ± 0.18
Conjunction	0.182 ± 0.19	0.274 ± 0.22	0.259 ± 0.12	0.056 ± 0.19	0.205 ± 0.12	0.195 ± 0.17
Coordination scope	0.048 ± 0.07	0.088 ± 0.13	−0.125 ± 0.15	0.136 ± 0.07	0.019 ± 0.10	0.033 ± 0.10
Core args	0.246 ± 0.02	0.140 ± 0.11	0.270 ± 0.15	0.168 ± 0.14	0.388 ± 0.05	0.242 ± 0.09
Datives	0.388 ± 0.12	0.337 ± 0.34	0.362 ± 0.29	0.268 ± 0.23	0.012 ± 0.21	0.273 ± 0.24
Disjunction	−0.188 ± 0.11	0.116 ± 0.03	0.074 ± 0.22	0.059 ± 0.11	−0.012 ± 0.21	0.010 ± 0.14
Double negation	0.142 ± 0.08	−0.001 ± 0.07	−0.235 ± 0.23	0.041 ± 0.21	0.076 ± 0.13	0.005 ± 0.15
Downward monotone	0.135 ± 0.07	0.031 ± 0.11	0.012 ± 0.12	0.129 ± 0.09	0.066 ± 0.07	0.074 ± 0.09
Ellipsis/implicits	0.454 ± 0.03	0.113 ± 0.18	0.031 ± 0.14	0.028 ± 0.10	0.180 ± 0.22	0.161 ± 0.13
Existential	0.221 ± 0.09	0.355 ± 0.06	0.249 ± 0.21	0.122 ± 0.16	0.333 ± 0.19	0.256 ± 0.14
Factivity	0.286 ± 0.05	0.139 ± 0.09	0.150 ± 0.10	0.247 ± 0.04	0.197 ± 0.05	0.204 ± 0.07
Genitives/partitives	0.090 ± 0.20	0.085 ± 0.22	−0.027 ± 031	0.168 ± 0.17	0.462 ± 0.26	0.156 ± 0.23
Intersectivity	0.252 ± 0.08	0.158 ± 0.02	0.025 ± 0.14	0.258 ± 0.08	0.102 ± 0.10	0.159 ± 0.09
Intervals/numbers	−0.177 ± 0.10	−0.050 ± 0.10	−0.075 ± 0.10	−0.060 ± 0.12	−0.132 ± 0.07	−0.099 ± 0.10
Lexical entailment	0.048 ± 0.06	0.077 ± 0.08	0.048 ± 0.10	0.111 ± 0.12	0.076 ± 0.11	0.072 ± 0.09
Morphological negation	0.061 ± 0.09	0.126 ± 0.09	0.238 ± 0.08	0.207 ± 0.18	0.09 ± 0.09	0.144 ± 0.11
Named entities	0.123 ± 0.09	−0.072 ± 0.18	0.152 ± 0.22	0.160 ± 0.14	0.143 ± 0.14	0.101 ± 0.15
Negation	0.088 ± 0.01	0.057 ± 0.13	−0.027 ± 0.16	0.093 ± 0.04	0.045 ± 0.10	0.051 ± 0.09
Nominalization	0.142 ± 0.21	0.002 ± 0.01	0.117 ± 0.11	0.295 ± 0.22	−0.006 ± 0.15	0.110 ± 0.14
Non-monotone	0.076 ± 0.07	−0.102 ± 0.26	0.122 ± 0.13	0.120 ± 0.14	0.063 ± 0.23	0.056 ± 0.17
Prepositional phrases	0.402 ± 0.08	0.293 ± 0.09	0.255 ± 0.15	0.403 ± 0.10	0.380 ± 0.12	0.347 ± 0.11
Quantifiers	0.056 ± 0.07	0.252 ± 0.17	0.124 ± 0.11	0.222 ± 0.19	0.242 ± 0.14	0.179 ± 0.14
Redundancy	0.00 ± 0.00	0.075 ± 0.24	0.047 ± 0.16	0.503 ± 0.23	0.287 ± 0.32	0.182 ± 0.19
Relative clauses	0.081 ± 0.05	0.248 ± 0.15	0.303 ± 0.21	0.247 ± 0.20	0.169 ± 0.17	0.210 ± 0.15

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

Table 6. Continued.

Feature	English	French	German	Russian	Swedish	Average
Restrictivity	-0.094 ± 0.20	-0.016 ± 0.04	0.112 ± 0.22	-0.048 ± 0.30	0.198 ± 0.23	0.030 ± 0.20
Symmetry/collectivity	-0.024 ± 0.05	-0.143 ± 0.07	-0.009 ± 0.18	-0.202 ± 0.12	0.428 ± 0.17	0.010 ± 0.12
Temporal	0.002 ± 0.22	-0.007 ± 0.11	0.063 ± 0.13	0.093 ± 0.21	-0.051 ± 0.05	0.020 ± 0.14
Universal	0.396 ± 0.19	0.725 ± 0.15	0.495 ± 0.14	0.343 ± 0.23	0.345 ± 0.10	0.461 ± 0.16
Upward monotone	0.292 ± 0.09	0.300 ± 0.13	0.196 ± 0.19	0.345 ± 0.19	0.395 ± 0.07	0.306 ± 0.14
World knowledge	0.090 ± 0.12	0.090 ± 0.11	0.077 ± 0.10	0.035 ± 0.09	0.128 ± 0.10	0.084 ± 0.10

Figures 6–9 show the results of the diagnostic evaluation of the mBERT model fine-tuned with multiple random seeds on the corresponding RTE dataset for each language (see Section 5.2).

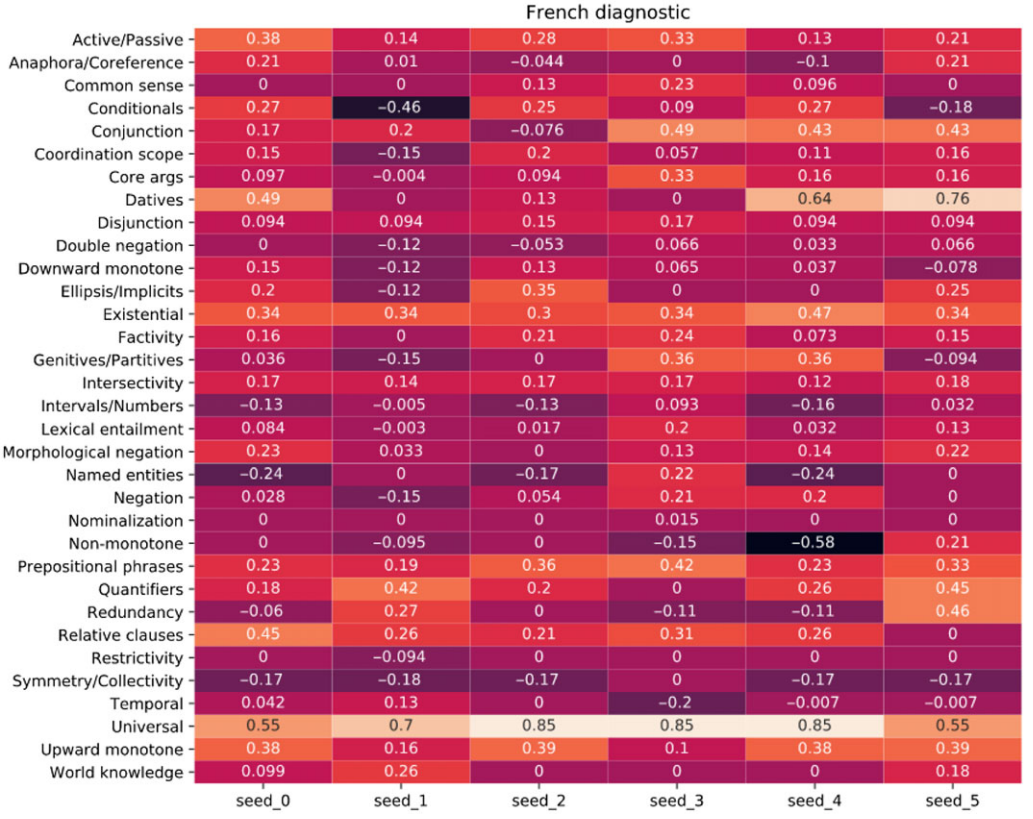

Figure 6. MCC scores on the French diagnostic dataset for mBERT fine-tuned with multiple random seeds.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

Figure 7. MCC scores on the German diagnostic dataset for mBERT fine-tuned with multiple random seeds.

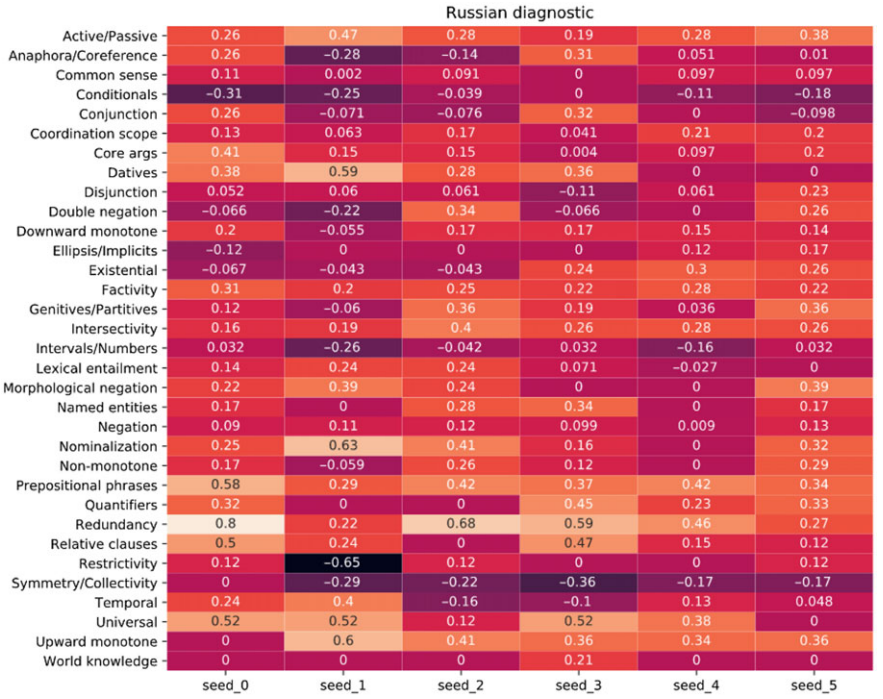

Figure 8. MCC scores on the Russian diagnostic dataset for mBERT fine-tuned with multiple random seeds.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

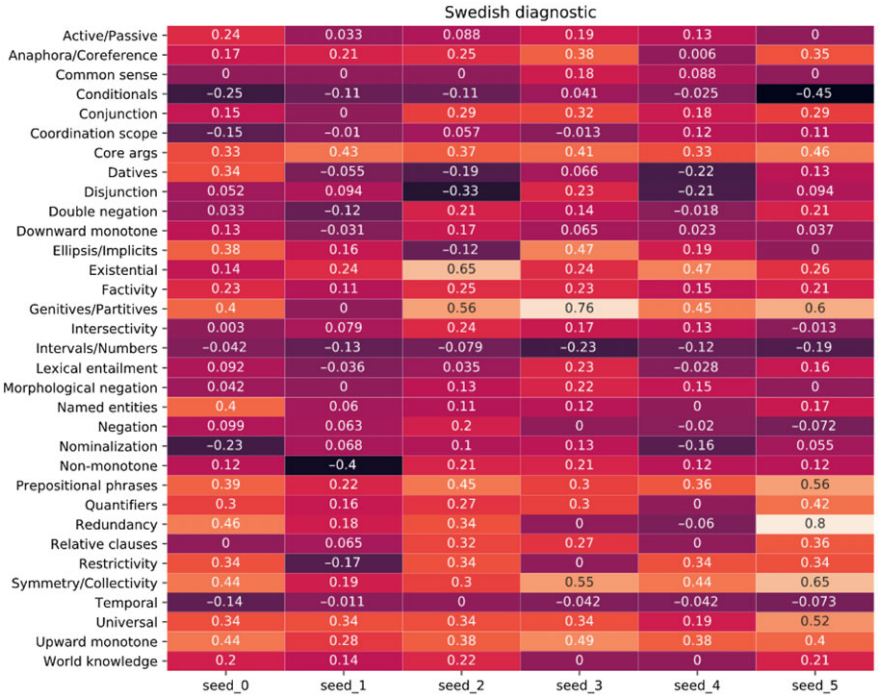

Figure 9. MCC scores on the Swedish diagnostic dataset for mBERT fine-tuned with multiple random seeds.

8.2 Fine-tuning stability and dataset size

Figure 10 depicts the results of the language-wise diagnostic evaluation of mBERT when fine-tuned on combined RTE and MNLI training samples. Comparing the heat map with that of Figure 1 (see Section 5.2), we observe that MCC scores for some categories have greatly improved for all languages (*Conjunction*, *Coordination scope*, *Core args*, *Genitives/Partitives*, *Prepositional phrases*, and *Universal*), while logic categories negatively correlate with the model predictions (*Disjunction*, *Downward monotone*, and *Intervals/Numbers*). Figures 11–15 show seed-wise diagnostic evaluation of the mBERT model when fine-tuned on combined RTE and MNLI training datasets with multiple random seeds.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

Figure 10. Language-wise diagnostic evaluation of mBERT when fine-tuned on combined RTE and & MNLI training datasets.

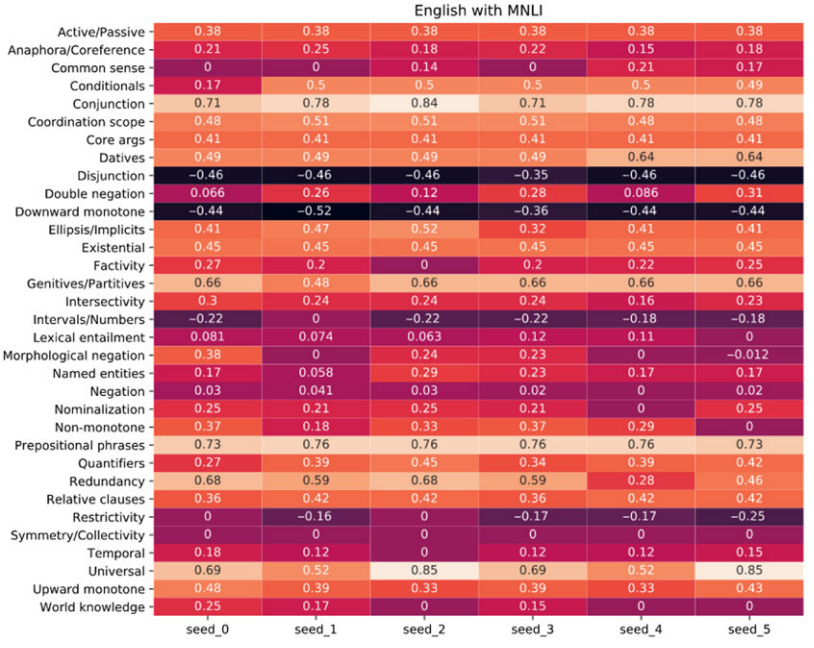

Figure 11. mBERT's seed-wise English diagnostic evaluation when fine-tuned on combined RTE and MNLI training datasets.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

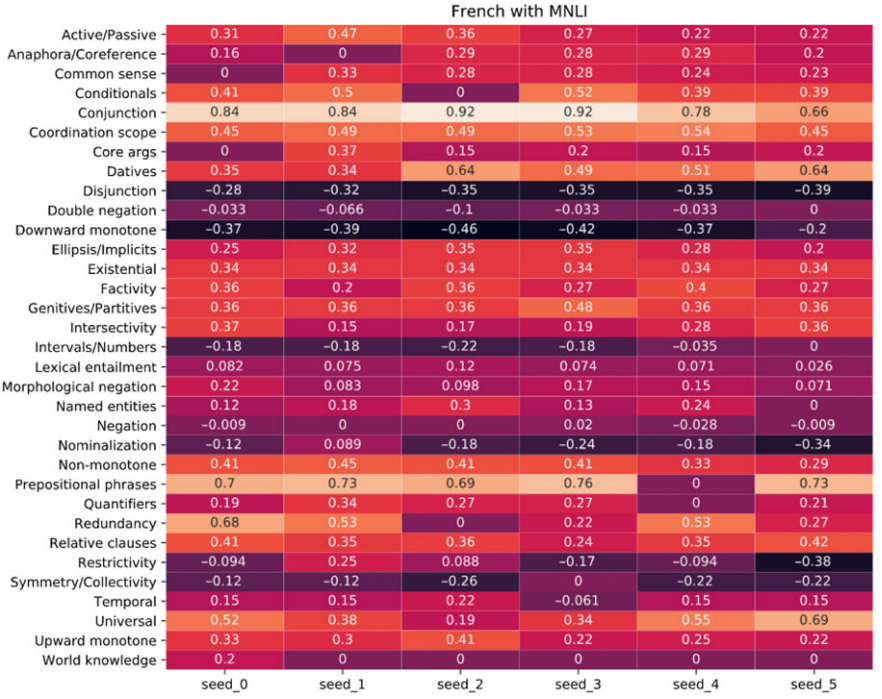

Figure 12. mBERT's seed-wise French diagnostic evaluation when fine-tuned on combined RTE and MNLI training datasets.

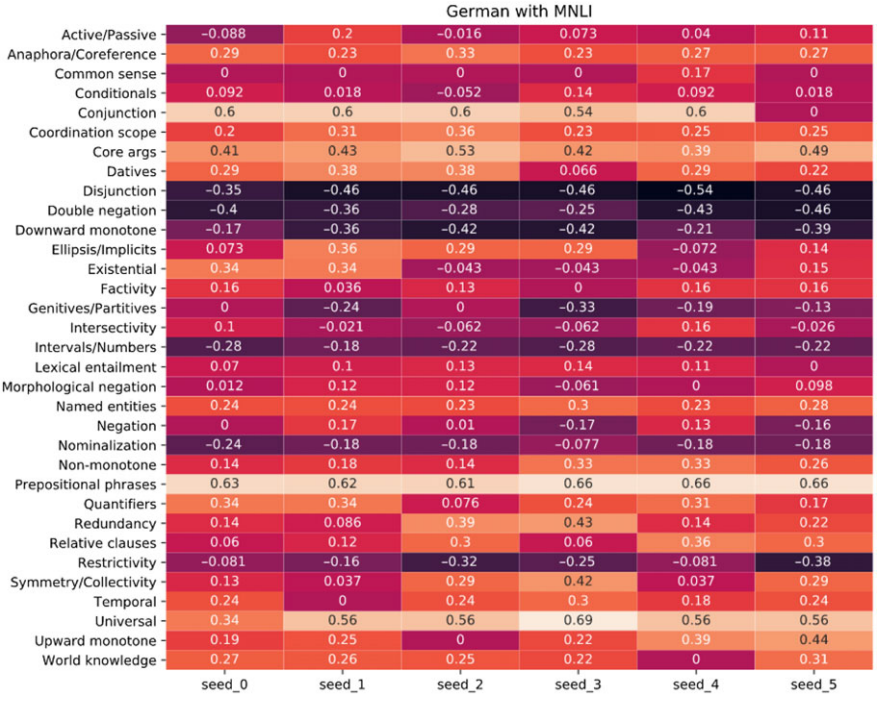

Figure 13. mBERT's seed-wise German diagnostic evaluation when fine-tuned on combined RTE and MNLI training datasets.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

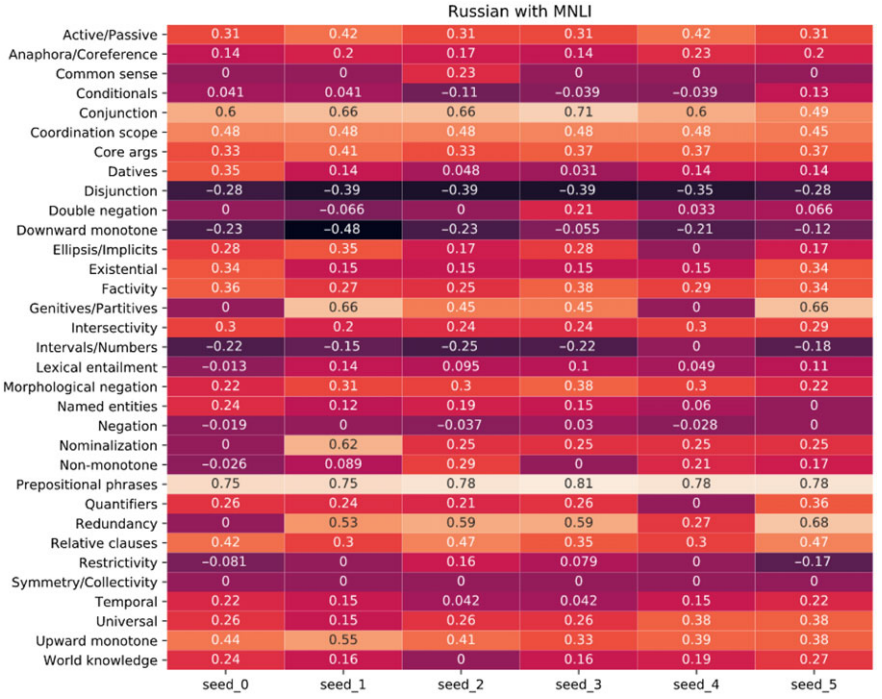

Figure 14. mBERT's seed-wise Russian diagnostic evaluation when fine-tuned on combined RTE and MNLI training datasets.

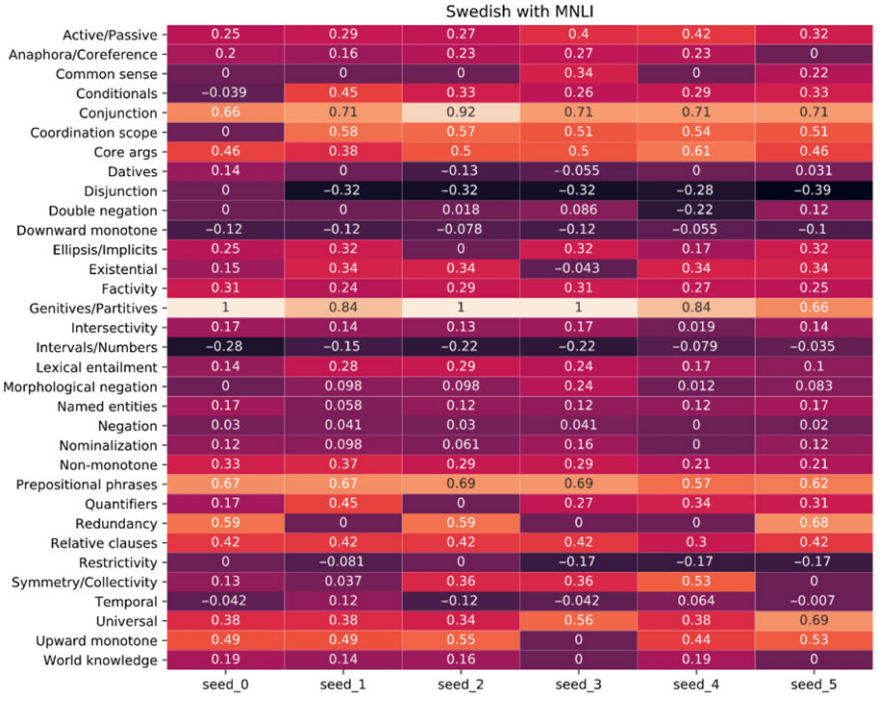

Figure 15. mBERT's seed-wise Swedish diagnostic evaluation when fine-tuned on combined RTE and MNLI training datasets.

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

8.3 Automatic annotation of diagnostic features

Table 7. Distribution of 15 diagnostic features in the RTE training datasets for English and Russian, and in the 10k MNLI subset according to the automatic annotation pipeline.

Feature	TERRa	RTE	MNLI 10k
Double negation	0.015	0.009	0.010
Negation	0.467	0.086	0.134
Intervals/numbers	0.155	0.517	0.149
Temporal	0.066	0.145	0.047
Conditionals	0.043	0.043	0.067
Conjunction	0.528	0.601	0.467
Disjunction	0.034	0.072	0.083
Prepositional phrases	0.963	0.993	0.925
Anaphora/coreference	0.676	0.547	0.679
Relative clauses	0.203	0.451	0.340
Datives	0.568	0.100	0.099
Quantifiers	0.187	0.143	0.217
Factivity	0.094	0.059	0.082
Morphological negation	0.394	0.378	0.255
Named entities	0.322	0.893	0.461

8.4 Coarse-grained probing analysis

A prominent methodology to explore the inner workings of pretrained LMs is to train a lightweight classifier over features produced by them to predict a linguistic property. During the probing procedure, the hidden representations produced by the model are taken from various layers of the transformer, and then a simple classifier is trained to predict a linguistic feature based on the given supervision (e.g., whether a particular category is present in a sentence or not). The underlying assumption is that if the classifier can predict the property, then the representations implicitly encode the linguistic knowledge.

We apply the annotation procedure (see Section 5.4) to create a set of three binary classification tasks for English and Russian that correspond to the coarse-grained diagnostic categories of *Logic*, *Lexical Semantics*, and *Predicate-Argument structure*. The task is to identify if a particular category is present in a given pair of sentences. We follow the SentEval probing methodology (Conneau et al., 2018a) to train a linear classifier using cross-entropy loss, optimized with Adam (Kingma and Ba 2014). The classifier is trained on the corresponding annotated RTE dataset's concatenated train and validation sets. We tune the L2-regularization parameter $\in [0.1, ..., 1e^{-5}]$ on the RTE test set and evaluated performance on the diagnostic set using accuracy score. The input to the classifier is a concatenation of the mean-pooled intermediate representations of each sentence in a given pair. We probe a pretrained mBERT model as a reference, and six mBERT models fine-tuned on the RTE task with multiple random seeds $\in [0; 5]$ (see Section 5.2).

We now provide a brief description of the probing results. The overall pattern is that the probing trajectories across the models are more consistent for English than Russian. Specifically, the

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

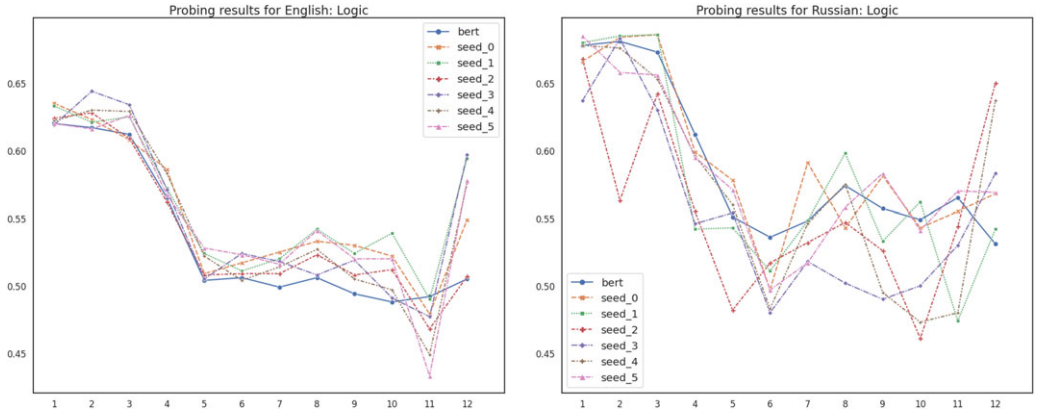

Figure 16. Probing results for the category of Logic. X-axis is the layer number, while Y-axis refers to the classifier performance (accuracy score).

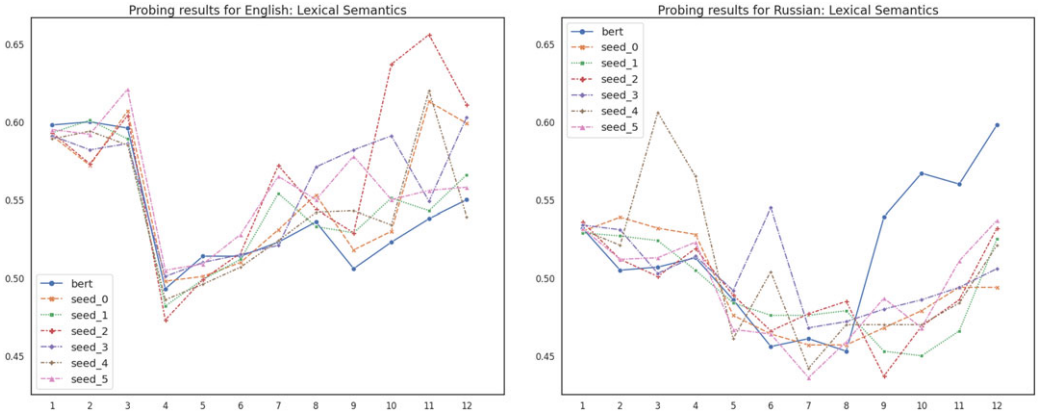

Figure 17. Probing results for the category of lexical semantics. X-axis is the layer number, while Y-axis refers to the classifier performance (accuracy score).

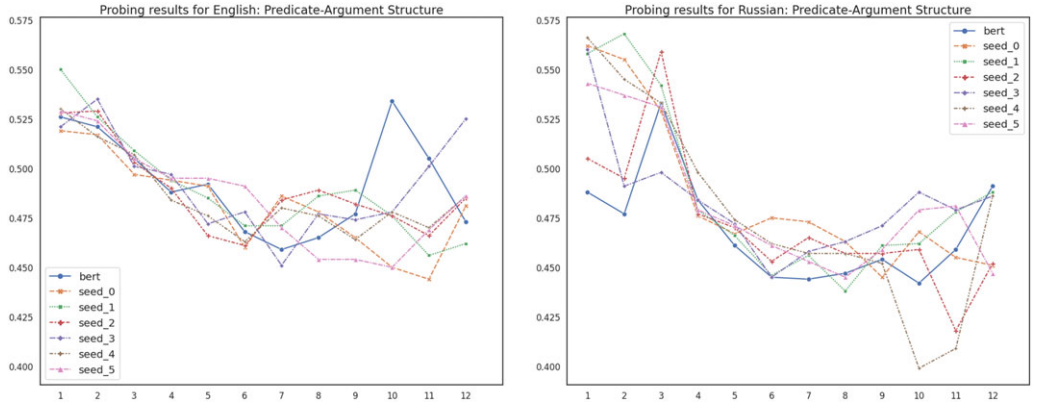

Figure 18. Probing results for the category of predicate-argument structure. X-axis is the layer number, while Y-axis refers to the classifier performance (accuracy score).

https://doi.org/10.1017/S1351324922000225 Published online by Cambridge University Press

linguistic properties tend to be more localized in the lower layers than in the higher ones, meaning that the latter is more affected by the fine-tuning (Wu et al., 2020). Note that the lower and middle layers of the models for Russian are less similar, which is demonstrated by sharp increases and decreases in the probe performance. Besides, the fine-tuning effect differs across the tasks, for example, leading to better performance over the *Lexical Semantics* task for English, and vice versa for Russian (see Figure 17). This can be interpreted as follows: the fine-tuning unpredictably causes the model either to "forget" about a particular knowledge or to "acquire" the knowledge of low certainty, shown over several RS models for both English and Russian. Despite the varying trajectories, the performance results remained similar for both languages, ranging from being close to or below random choice (see Figures 17 and 18) to becoming more confident in the lower layers on the *Logic* tasks (see Figure 16), with overall quality around 65% accuracy score.

Cite this article: Tikhonova M, Mikhailov V, Pisarevskaya D, Malykh V and Shavrina T (2023). Ad astra or astray: Exploring linguistic knowledge of multilingual BERT through NLI task. *Natural Language Engineering* **29**, 554–583. https://doi.org/10.1017/S1351324922000225