Diacritics Restoration using BERT with Analysis on Czech language

Jakub Náplava, Milan Straka, Jana Straková

Institute of Formal and Applied Linguistics Charles University, Czech Republic Faculty of Mathematics and Physics

Abstract

We propose a new architecture for diacritics restoration based on contextualized embeddings, namely BERT, and we evaluate it on 12 languages with diacritics. Furthermore, we conduct a detailed error analysis on Czech, a morphologically rich language with a high level of diacritization. Notably, we manually annotate all mispredictions, showing that roughly 44% of them are actually not errors, but either plausible variants (19%), or the system corrections of erroneous data (25%). Finally, we categorize the real errors in detail. We release the code at https://github.com/ufal/bert-diacritics-restoration.

1. Introduction

Diacritics Restoration, also known as Diacritics Generation or Accent Restoration, is a task of correctly restoring diacritics in a text without any diacritics. Its main difficulty stems from ambiguity where context needs to be taken into account to select the most appropriate word variant, because diacritization removal creates new groups of homonymy.

Current state-of-the-art algorithms for diacritics restoration are mostly based on either recurrent neural networks combined with an external language model (Náplava et al., 2018; AlKhamissi et al., 2020) or Transformer (Mubarak et al., 2019). Recently, BERT (Devlin et al., 2019) was shown to outperform many models on many tasks while being much faster due to the fact that it uses simple parallelizable classification head instead of a slow auto-regressive approach.

In this work, we first describe a model for diacritics restoration based on BERT and evaluate it on multilingual dataset comprising of 12 languages (Náplava et al., 2018).

We show that the proposed model outperforms the previous state-of-the-art system (Náplava et al., 2018) in 9 languages significantly.

We further provide an extensive analysis of our model performance in Czech, a language with rich morphology and a high level of diacritization. In addition to clean data from Wikipedia (Náplava et al., 2018), the model was evaluated on data collected from other domains, including noisy data, and we show that stable performance holds even if the text contains spelling and other grammatical errors.

Sometimes, multiple plausible diacritization variants are possible, while only one gold reference exists, which comes from the original text before diacritization was automatically stripped to create test data. To assess the extent of these cases, we employed annotators to manually annotate all mispredictions and we found that 19% of errors are plausible variants and 25% of errors are system corrections of errors in data.

Finally, we further analyse the remaining errors by analysing characteristics of plausible variants.

2. Related Work

Diacritics Restoration is an active area of research in many languages: Vietnamese (Nga et al., 2019), Romanian (Nuţu et al., 2019), Czech (Náplava et al., 2018), Turkish (Adali and Eryiğit, 2014), Arabic (Madhfar and Qamar, 2020; AlKhamissi et al., 2020) and many others.

There are three main architectures currently used in diacritics restoration: convolutional neural networks (Alqahtani et al., 2019), recurrent neural networks often combined with an external language model (Belinkov and Glass, 2015; Náplava et al., 2018; AlKhamissi et al., 2020) and Transformer-based models (Orife, 2018; Mubarak et al., 2019). The convolutional neural networks are fast to train and also to infer. However, compared to the recurrent and Transformer-based architectures, they do generally achieve slightly worse results due to the fact that they model long-range dependencies worse. On the other hand, recurrent- and Transformer-based architectures are much slower.

Recently, the BERT model (Devlin et al., 2019) comprising of self-attention layers, was proposed and shown to reach remarkable results on a variety of tasks. As it uses no recurrent layers, its inference time is much shorter. We expect BERT to significantly improve the performance over current state-of-the-art diacritization architectures.

3. Model Architecture

The core of our system is a pre-trained multilingual BERT model that uses self-attention layers to create contextualized embeddings for tokenized text without diacritics. The contextual embeddings are fed into a fully-connected feed-forward neural network followed by a softmax layer. This outputs a vector with a distribution over

0.5%	<KEEP>
...	
79%	1:ACUTE;3:CARON
...	
0%	0:RING ABOVE;3:CARON

FFNN + softmax

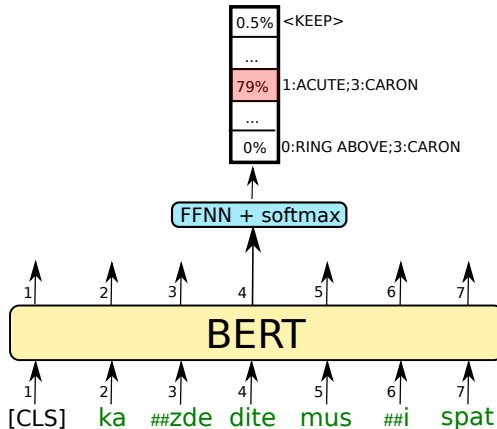BERT

[CLS] ka ##zde dite mus ##i spat

Figure 1. Model architecture. Text without diacritics, tokenized into subwords, is fed to BERT and for each of its outputs, fully-connected network followed by softmax is applied to obtain the most probable instruction for diacritization. ##-prefixes of some subwords are added by the BERT tokenizer.

a set of instructions that define diacritization operation over individual characters of each input token. We select the instruction with maximum probability. The model is illustrated in Figure 1.

3.1. Diacritization Instruction Set

To decrease the size of the final softmax layer, the output labels are not the diacritized variants of input subwords, as one would expect, but they are a set of instructions that provide prescription on how to restore diacritics. Specifically, one such instruction consists of index-diacritical mark tuples that define on what index of input subword a particular diacritical mark should be added.

An example of a diacritization instructions set can be seen in Figure 2. Given an input subword *dite (dítě)*, with four characters indexed from 0 to 3, the appropriate diacritization instruction is *1:ACUTE;3:CARON*, in which acute is to be added to *i* and caron is to be added to *e* resulting in a properly diacritized word *dítě*. Obviously, the network can choose to leave the (sub)word unchanged, for which a special instruction *<KEEP>* is reserved. Should the network accidentally select an impossible instruction, no operation is carried out and the input (sub)word is also left unchanged.

To construct the set of possible diacritization instructions, we tokenize the undiacritized text of the particular training set and align each input token to the corresponding token in the diacritized text variant. The diacritical mark in each instruction is obtained from the Unicode name of the diacritized character. We keep only those

3

input	instruction	result	note
dite	1:CARON;3:ACUTE	dítě	optimal instruction
dite	1:CARON	díte	
dite	3:ACUTE	ditě	
dite	<KEEP>	dite	no change
dite	2:RING ABOVE	dite	impossible instruction ignored

Figure 2. Diacritization instructions examples for input "dite (dítě)" with 4 characters, indexed from 0 to 3. Index-Instruction tuples generate diacritics for given input.

instructions that occurred at least twice in a training set to filter out extremely rare instructions that originate for example from foreign words or bad spelling.

3.2. Training Details

We train both the fully-connected network and BERT with AdamW optimizer which minimizes the negative log-likelihood. The learning rate linearly increases from 0 to 5e-5 over the first 10000 steps and then remains the same. We use HuggingFace implementation of *BertForTokenClassification* and initialize *BERT-base* values from *bert-base-multilingual-uncased* model.

We use the batch size of 2048 sentences and clip each training sentence on 128 tokens. We train each model for circa 14 days on Nvidia P5000 GPU and select the best checkpoint according to development set.

4. Automatic Evaluation on Diacritization Corpus with 12 Languages

We evaluate our approach on the dataset of Náplava et al. (2018). This dataset contains training and evaluation data for 12 languages: Vietnamese, Romanian, Latvian, Czech, Polish, Slovak, Irish, Hungarian, French, Turkish, Spanish and Croatian.

We evaluate the model performance using a standard metric, the *alpha-word accuracy*. This metric omits words composed of non-alphabetical characters (e.g., punctuation).

For each language, we compute an independent set of operations and train a separate model. We use the concatenation of the Wiki and the Web training data of (Náplava et al., 2018) both for computing a set of instructions and also as the training data for our model.[1] The size of each instruction set and our results in comparison

---

[1]In Romanian Web data, ş (LATIN SMALL LETTER S WITH CEDILLA) is for historical reasons often used instead of ș (LATIN SMALL LETTER S WITH COMMA BELOW) and similarly ţ (LATIN SMALL LETTER T WITH CEDILLA) is often used instead of ț (LATIN SMALL LETTER T WITH COMMA BELOW). We replace the occurrences of the previously-used characters (the former ones) with their standard versions (the latter ones).

Language	Instruction Set Size	Náplava et al. (2018)	Ours	Error Reduction
Czech	1005	99.06	**99.22 ±0.046**	17 %
Vietnamese	2018	97.73	**98.53 ±0.037**	35 %
Latvian	720	97.49	**98.63 ±0.045**	45 %
Polish	1005	99.55	**99.66 ±0.041**	24 %
Slovak	785	99.09	**99.32 ±0.030**	25 %
French	681	**99.71**	**99.71 ±0.016**	0 %
Irish	189	98.71	**98.88 ±0.040**	13 %
Spanish	492	**99.65**	99.62 ±0.018	− 9 %
Croatian	541	99.67	**99.73 ±0.018**	18 %
Hungarian	767	99.29	**99.41 ±0.038**	17 %
Turkish	1005	**99.28**	98.95 ±0.046	− 46 %
Romanian	1677	98.37	**98.64 ±0.056**	17 %

Table 1. Comparison of alpha-word accuracy of our model including 95% confidential intervals to previous state-of-the-art on 12 languages.

with the previous state-of-the-art-results of Náplava et al. (2018) are presented in Table 1. Apart for alpha-word accuracy itself, we also report 95% confidential intervals computed using bootstrap resampling method.

On 9 of 12 languages, our approach significantly outperforms previous state-of-the-art combined recurrent neural networks with an external language model. The most significant improvements are achieved on Vietnamese and Latvian.

5. Detailed Analysis on Czech

We further provide a detailed analysis of our model performance in Czech, a language with rich morphology and a high diacritization level: Of the 26 English alphabet letters, a half of them can have one or two kinds of diacritization marks (Zeman, 2016). Czech is also the 4-th most diacritized language of the 12 languages found in the diacritization corpus of Náplava et al. (2018).

Particularly, we are interested in the three following questions:

- How would our system perform outside the very clean Wiki domain? (Section 5.1)
- Is it possible that some of the labeled mispredictions are actually plausible variants? (Section 5.2)
- Is there an observable characteristics in the real errors made by the system? (Section 5.3)

5

Domain	Sentences	Words	Evaluated Words
Natives Formal	1 743	19 973	19 138
Natives Informal	7 223	99 352	86 720
Romi	1 490	15 971	13 080
Second Learners	5 117	63 859	50 630

Table 2. Basic statistics of new data for testing diacritics restoration in Czech.

5.1. Additional Domains

The testing dataset of Náplava et al. (2018) is composed of clean sentences originating from Wikipedia. It is, however, a well-known fact that the performance of the (deep neural) models may deteriorate substantially when the input domain is changed (Belinkov and Bisk, 2017; Rychalska et al., 2019). To test our system in other, more challenging domains, we used data from a new Czech dataset (unpublished, in annotation process) for grammatical-error-correction that contains data collected from 4 sources:

- Natives Formal – Essays of elementary school Czech pupils (decent Czech proficiency)
- Natives Informal – texts collected from web discussions
- Second Learners – essays of Czech second learners
- Romi – texts of Czech pupils with Romani ethnolect (low Czech proficiency)

The dataset covers a wide range of Czech domains. It contains texts annotated in M2 format, a standard annotation format for grammar-error-correction corpora. In this format, each document contains original sentences with potential errors (e.g. spelling, grammatical or errors in diacritics) and a set of annotations describing what operations should be performed in order to fix each error.

To create target data for diacritics restoration, we apply all correcting edits that fix errors in diacritics and casing. We leave other errors intact, but do not evaluate on words that contain these errors, because they are not directly relevant to diacritics and in many cases, the errors are so severe that evaluation would be controversial. To rule out such words, we create a binary mask that distinguishes between evaluated and omitted words. Although the severely perturbed words are omitted from evaluation, they still remain in the sentence context and may still confuse the diacritization system, making the task potentially more difficult. See examples of such misleading sentence contexts in Figure 3.

The basic statistics of the new dataset are presented in Table 2. We display the number of sentences, the number of all words and the number of evaluated (unmasked) words. Compared to the Wikipedia dataset (Náplava et al., 2018), our new dataset has half the number of sentences and one third of its number of words.

> Potřebujeme nové idea i <u>novych</u> **lidi**/lidí* , ktery je přinesou .
>
> Na ulicích vidíme často některý lidi , kteří nosí **barevné**/barevně* <u>oblečeny</u> , které jsou snad hezké , ale určitě nejsou elegantní .

> *English translation (without ambiguities)*
>
> *We need <u>new</u> ideas and also **people** to come up with them.*
>
> *In the streets, we can see some people wearing **colourful** <u>clothes</u>, which may be nice but certainly not elegant.*

Figure 3. Examples of misleading contexts in noisy texts. Correct diacritization (bold) can only be achieved by grammar corrections of the surrounding words (underlined).

We evaluate our model on all the above introduced Czech domains and present the results in Table 3. Despite our initial concern that the model would perform worse on these domains due to the noisy nature of the data, the results show that the model performance remains roughly stable on all domains. We suppose that although the writers produced quite noisy texts, they at the same time avoided foreign words that are generally harder to correctly diacritize.

5.2. Error Annotation

Clearly, removing diacritics creates new groups of homonyms (*dal/dál*, *krize/kříže*). In most cases, the correct diacritization variant can be inferred by a method which takes the sentence context into consideration. However, there are cases, in which more plausible variants are available, e.g., *šachu/šachů*, *pradlena/přadlena*, *podána/podaná*, as illustrated in Figure 4. Furthermore, some variants can only be disambiguated in the context of the whole document, such as in: *K nejvýznamnějším patří zmiňované vily/víly.* (more examples in Figure 6), not to mention other examples that can be only disambiguated by real-world knowledge such as in *Povrch satelitu/satelitů Země už zkoumalo několik sond.*

However, all our evaluation data are limited only to a single gold reference for each word without diacritics, given by the fact that the gold reference comes from the original text with diacritics. To explore both phenomena among the mispredictions, we hired annotators to examine: a) whether a word is correctly diacritized given the context of current sentence; and b) whether it is correct given a context of two previous sentences, current sentence and two following sentences (thus ruling out the words with even longer document dependencies).

While the evaluation of the clear Wiki data (Náplava et al., 2018) is straightforward, some of our newly introduced noisy data may become controversial to evaluate due

Nebo záměna kapitol a jejich časová posloupnost v knize je pak ve filmu **podána/podaná** rozdílně .

Hraní **šachu/šachů** , ale především karetních her , kritizoval také Petr Chelčický .

Jeho matka byla **přadlena/pradlena** , která ke sklonku života propadla alkoholu .

Hororová hudba slouží především pro dokreslení **filmů/filmu** .

English translation

*The chapters and their chronological order in the book are then **presented/given** differently in the film.*

*Playing **a game of chess/games of chess** , but especially card games was criticized by Petr Chelčický .*

*His mother was a **washerwoman/laundress** who fell into alcoholism towards the end of her life .*

*Horror music is mainly used to complete **a movie/movies** .*

Figure 4. Examples of ambiguities, each illustrating two diacritization variants (bold), both valid in a given context.

to erroneous words. Therefore, such words were also marked by the annotators and subsequently removed from our analysis.

An example of a final annotation item presented to an annotator is illustrated in Figure 5.

To create the annotation items, we concatenated data from all domains, both the original Wikipedia data (Náplava et al., 2018) and other domains (Section 5.1) and we further considered those words in which the results of our system did not match target word. Before annotation, we automatically filtered out some cases:

- Predictions, in which the system and the target words are variants (as marked by MorphoDita (Straková et al., 2014)) were automatically marked correct.
- Predictions, in which the target word was marked as non-existing by MorphoDiTa, while the system word was marked as Czech, were considered dubious and removed from our analysis.

For the remaining 4702 words, two annotation items were created: one with the predicted word and one with the gold reference word in the position of the annotated *Current Word*. The annotation process took circa 70 hours.

The basic analysis of the annotated system errors is the following: There are 4702 wrongly diacritized words in the all our data concatenated. Annotations revealed that 960 of the mispredicted words contain a non-diacritical error and we do not consider

Předpřechozí věta	Popisujeme sítě , které nepoužívají sdílený přenosový prostředek .
Předchozí věta	Přenosové rychlosti se velmi liší podle typu sítě .
Začátek aktuální věty	Začínají na desítkách kilobitů za sekundu , ale dosahují i
Aktuální slovo	**rychlosti**
Konec aktuální věty	řádu několik gigabitů za sekundu .
Následující věta	Příkladem takové sítě může být internet .
Věta po následující větě	Mezi rozlehlé sítě patří :
Je správně vůči aktuální větě:	Ano
Je správně vůči cel. kontextu:	Ne
Obsahuje překlep:	Ne

	English translation
Before Previous Sentence:	*We describe networks that do not use a shared transmission medium .*
Previous sentence:	*Transmission speeds vary greatly depending on the type of network .*
Current Sentence Start:	*They start at tens of kilobits per second , but also reach*
Current Word:	***speeds***
Current Sentence End	*of the order of a few gigabits per second .*
Next Sentence:	*An example of such a network is the Internet.*
After Next Sentence:	*Large networks include :*
Is Correct w.r.t. Cur. Sentence:	*True*
Is Correct w.r.t. Whole Context:	*False*
Contains Spelling Typo:	*False*

Figure 5. Annotation item example. The annotator marks whether the word "rychlosti" is correct given a context of the current sentence, whether it is still correct in the context of two previous and two following sentences and whether it contains a typo.

them further, as mentioned above. The remaining 3742 mispredicted words can be categorized as follows:

- System correct, Gold correct: 19% (694 of 3742) – plausible variants
- System correct, Gold wrong: 25% (964 of 3742) – system corrects data error
- System wrong, Gold wrong 1% (31 of 3742) – uncorrected error in data
- System wrong, Gold correct 55% (2 084 of 3742) – real errors

Interestingly, the annotations revealed that about 44% of errors are not errors at all. In 694 cases (19%) both the system word and the gold word are correct, which is justified by the plausible variants. In 964 cases (25%) the original gold annotation was wrong whereas the system annotation was correct, which means that the system effectively corrected some of the errors in the original data. The remaining 31 cases are for neither the system nor the gold word being correct. Finally, the annotations confirmed 2084 real system errors, which we postpone for a more detailed analysis in the following Section 5.3.

Plausible variants, which constitute 19% of the annotated errors, are the most interesting item. Please note that our criterion for plausible variant was strict: only

Domain	Original	Annotated	Annotated w/o annotated typos
Wiki	99.22	99.49	99.66
Natives Formal	99.50	99.75	99.75
Natives Informal	99.12	99.53	99.62
Romi	99.11	99.46	99.54
Second Learners	99.18	99.73	99.79

Table 3. Alpha-word accuracy of Czech model on 5 datasets from various domains.

cases ambiguous both in the sentence and document context were marked as plausible variants. Circa 72% percent of these words share a common lemma. As Table 4.a and Table 5.a show, singular/plural ambiguities by far most often arise in inanimate masc. genitive (*programu/programů*, *šachu/šachů*). Another common ambiguity is passive participle vs. adjective (*založena/založená*), generally known to be difficult for diacritization disambiguation (Zeman, 2016). More interesting examples are given in Table 4.a and Table 5.a.

To conclude, we use the collected annotations to refine our previous results, which we display in Table 3. When considering all annotated words, including those pre-processed with MorphoDiTa, we achieve 35% to 67% error reduction. When omitting words newly marked by human annotators as containing another (non-diacritical) error, the error rate gets additionally reduced by up to 33%.

5.3. Analysis of Real Errors

We follow with a morphological analysis of the remaining confirmed errors, which constitute 55% of the annotated mispredictions. To determine the morphological categories of the erroneously predicted words, we use UDPipe (Straka et al., 2019) to generate morphological annotations for all words in model hypotheses and gold sentences. We then inspect the most frequent confusions between the system and the gold morphological annotations of words, using the Universal POS tags and Universal features (Nivre et al., 2020).

The annotations confirmed an interesting discourse phenomenon: a word can be correctly diacritized in multiple ways given the context of its sentence, however only a single correct diacritization variant exists if a wider context is taken into account. There are 50 such annotated cases; two examples are displayed in Figure 6. Although this phenomenon is interesting from a discourse perspective, its low proportion to actual errors (50 of 2084) indicates that it is quite rare. This implies that training models on longer texts (we currently train our model on examples comprising maximally 128 subwords – see Section 3.2) does not promise potential for overall improvement.

Finally, we offer a categorization of such ambiguities by means of the Universal POS tags and Universal features (Nivre et al., 2020) in Table 4.b and Table 5.b, respectively.

The remaining errors are a mix of complicated disambiguation cases or rare named entities. The most frequent errors bear similarity to plausible variants (compare Table 5.a and Table 5.c), only with a different order of appearance. Unlike plausible variants (Table 5.a), most frequent mismatches occur already at the level of lemmas (*stát/stať*, *že/ze*, see Table 5.c). Second most frequent cases are rare named entities (*Sokrates/Sókratés*, *Aristoteles/Aristotelés*, *Diogenés/Díogenés*). Number is again often hard to disambiguate in inanimate masc. genitive (*milionu/milionů*, *reproduktoru/reproduktorů*, *dokumentu/dokumentů*), followed by fem. case (*ji/jí*, *ni/ní*, *zemi/zemí*).

6. Conclusion

We implemented a model for diacritics restoration based on BERT that outperforms previous state-of-the-art models. Further analysis on Czech data collected from additional, noisy domains shown that the model exhibits strong performance regardless the domain of the data.

We further annotated all reported mispredictions in Czech and found out that more than one correct variant is sometimes possible. Rarely, disambiguation on document level is necessary to distinguish between variants correct within the sentence context. We elaborated on these phenomena using morphological annotations and utilized them to further analyse real confirmed errors of the systems.

As for future work, we propose experimenting with a single joint model for a subset of languages, despite our initial unsuccessful attempts at training a single model for all languages, including an introduction of a larger XLM-Roberta model (Conneau et al., 2020).

Tento motiv může být ovlivněn sibiřským šamanismem a průvodce pak má funkci psychopompa .

Kromě bohů znali pohanští Slované i celou řadu <u>nižších bytostí</u> , nazývány byly většinou slovem běs či div , které souvisí s indickým déva .

K nejvýznamnějším patří zmiňované **víly/vily** .

V různých podáních existují <u>víly</u> lesní , vzdušné , horské a také <u>víly</u> zlé .

Existují další ženské bytosti jim podobné , patří mezi ně především <u>rusalky</u> , <u>divé ženy</u> nebo <u>divoženky</u> doprovázené <u>divými muži</u> .

Další <u>dokumenty</u> týkající se Jana Žižky z Kalichu jsou <u>dva listy</u> odeslané z kláštera ve Vilémově datované k 16. březnu a 1. dubnu 1423 .

Slepý vojevůdce <u>v nich</u> vyzývá své straníky z orebského svazu k poradě naplánované na 7. či 8. dubna do Německého Brodu .

Z **dopisů/dopisu** je patrné , že se pokoušel dokonaleji zorganizovat husitskou vojenskou moc , pro boj s domácím i zahraničním nepřítelem .

O čtrnáct dní později Žižka spolu s orebity vedl válku se spojenci krále Zikmunda , zejména na Bydžovsku s panem Čeňkem z Vartenberka .

Tohoto šlechtice s jeho leníky a spojenci porazil 20. nebo 23. dubna v bitvě u Hořic , načež dál pokračoval v plenění jeho zboží .

English translation

This motif can be influenced by Siberian shamanism , and the guide then has the function of a psychopomp .

Apart from the gods, the pagan Slavs knew a number of <u>lower beings</u> , mostly called Raver or Wonder , which is related to Indian deva .

Among the most important are the mentioned **fairies/villas.**

There are wood <u>fairies</u>, air <u>fairies</u> , mountain <u>fairies</u> , and also evil <u>fairies</u> in various forms .

There are other female beings similar to them , they include mainly <u>mermaids</u> , <u>wild women</u> or <u>witches</u> accompanied by wild men .

Other <u>documents</u> concerning Jan Žižka of the Kalich are <u>two letters</u> sent from the monastery in Vilémov dated March 16 and April 1 , 1423 .

<u>In them</u> , the blind military leader invites his party members from the Orebic Union to a meeting scheduled for April 7 or 8 in Německý Brod .

The **letter shows/letters show** *that he has tried to better organize Hussite military power , to fight both domestic and foreign enemies.*

Fourteen days later , Žižka , together with the Orebits , waged war with King Zikmund's allies , especially in the Bydžov region with Mr. Čeněk of Vartenberk .

He defeated this nobleman with his feoffees and allies on April 20 or 23 at the Battle of Hořice , after which he continued to plunder his goods .

Figure 6. Two examples of ambiguous diacritization determined by document context.

Type	Count	Examples
NOUN ↔ NOUN	406	program[uů], šach[uů], text[uů]
ADJ ↔ ADJ	162	znám[áa], založen[aá], schopn[ií]
ADV ↔ ADJ	59	stejn[ěé], krásn[ěé], běžn[ěé]
PROPN ↔ PROPN	31	Aristotel[eé]s, Sokrates/Sókratés, J[aá]n
VERB ↔ VERB	20	zamýšlím/zamyslím, odráží/odrazí, os[ií]dlují
ADJ ↔ VERB	3	vznikl[áa], rádi/radí, splaskl[áa]
NOUN ↔ ADJ	2	přesvědčen[ií], očištěn[ií]
ADJ ↔ NOUN	2	veden[ií], považován[ií]
DET ↔ DET	2	jej[ií]ch, svoj[ií]

(a) Plausible variants.

Type	Count	Examples
NOUN → NOUN	32	stát/stať, objekt[uů], pulsar[uů]
VERB → VERB	4	narazí/naráží, řekn[ěe]te, žij[ií]
DET → DET	3	jej[ií]ch
ADJ → ADV	3	současn[éě], pravé/právě, praktick[ýy]
ADJ → ADJ	2	znám[áa], žádanou/zadanou
ADV → ADJ	2	stejn[ě/é]
NOUN → VERB	1	mysl[ií]

(b) Disambiguation from document context.

Type	Count	Examples
NOUN → NOUN	1596	stát/stať, lid[íi], program[uů]
PROPN → PROPN	587	Aristotel[eé]s, Sokrates/Sókratés, Kast[ií]lie
ADJ → ADJ	521	znám[aá], založen[aá], říd[ií]cí
VERB → VERB	193	m[ůu]že, M[aá]m, m[aá]
ADJ → ADV	134	krásn[éě], hezk[ýy], dobré/dobře
PRON → PRON	129	j[ií], n[ií], n[ií]ž
ADV → ADJ	112	stejn[ěé], pěkn[ěé], Obvykl[eé]
DET → DET	59	jej[ií]ch, svoj[ií], naš[ií]
NOUN → ADJ	47	mobiln[ií], brány/braný, češka/česká

(c) Real errors.

Table 4. Error categorization with universal POS. The context-dependent morphological annotations were obtained automatically using UDPipe.

Type	Count	Examples
Number	325	program[uů], šach[uů], objekt[uů]
Passive participle / adjective + more features	116	založen[aá], vzdálen[aá], nazýván[aá]
Lemma	82	l[eé]ty, mas[ií]vu, p[ée]rových
Adj ↔ Adv	59	stejn[éě], krásn[éě]
Variant + more features	31	znám[áa], schopn[ií], spokojen[ií]
Case	25	dr[aá]hami, dr[aá]hách, č[aá]rou
Lemma + more features	21	zamýšlím/zamyslím, ná[sš], pacht[uů]
Lemma, NameType	20	Aristotel[eé]s, Sokrates/Sókratés, [Íí]lias
Case, Number	8	boh[ůu], násobk[uů], funkc[ií]
Number, Person	5	považuj[ií], věnuj[ií], kupuj[ií]

(a) Plausible variants.

Type	Count	Examples
Lemma + more features	15	stát/stať, tvář/tvar, pravé/právě
Number	15	objekt[ůu], pulsar[uů], muzikál[ůu]
Lemma	6	řazení/ražení, v[ií]ly
Adj ↔ Adv	4	stejn[éě], současn[éě], praktick[ýy]
Case, Gender, Number	3	jej[ií]ch
Number, Person	2	narazí/naráží

(b) Disambiguation from document context.

Type	Count	Examples
Lemma + more features	924	stát/stať, [čc], [žz]e
Lemma, named entity + more features	382	D[ií]ogenés, Hal/Ħal, Dvořák/Dvorak
Number	226	milion[uů], reproduktor[ůu], dokument[ůu]
Case	149	j[ií], n[ií], zem[ií]
Adj ↔ Adv	132	pěkn[éě], česk[ýy], současn[éě]
Passive participle / adjective + more features	37	spojen[aá], pojmenovan[áa], prodaný/prodány
Case, Number	27	referent[uů], Dvořák[ůu], akademi[ií]
Case, Gender, Number	16	jej[ií]ch, j[ií]m
Number, Person	15	píš[ií], pracuj[ií], žij[ií]
Variant + more features	8	znám[áa], schopn[áa], hodn[áa]

(c) Real errors.

Table 5. Error categorization with extended Universal Features. The first column (Type) is the (primary) difference between the context-dependent feature sets of the system word and the gold word.

Bibliography

Adali, Kübra and Gülşen Eryiğit. Vowel and diacritic restoration for social media texts. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, pages 53–61, 2014. doi: 10.3115/v1/W14-1307.

AlKhamissi, Badr, Muhammad N ElNokrashy, and Mohamed Gabr. Deep Diacritization: Efficient Hierarchical Recurrence for Improved Arabic Diacritization. *arXiv preprint arXiv:2011.00538*, 2020.

Alqahtani, Sawsan, Ajay Mishra, and Mona Diab. Efficient Convolutional Neural Networks for Diacritic Restoration. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1442–1448, 2019. doi: 10.18653/v1/D19-1151.

Belinkov, Yonatan and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*, 2017.

Belinkov, Yonatan and James Glass. Arabic diacritization with recurrent neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2281–2285, 2015. doi: 10.18653/v1/D15-1274.

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

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

Madhfar, Mokhtar and Ali Mustafa Qamar. Effective Deep Learning Models for Automatic Diacritization of Arabic Text. *IEEE Access*, 2020.

Mubarak, Hamdy, Ahmed Abdelali, Hassan Sajjad, Younes Samih, and Kareem Darwish. Highly effective Arabic diacritization using sequence to sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2390–2395, 2019.

Náplava, Jakub, Milan Straka, Pavel Straňák, and Jan Hajic. Diacritics restoration using neural networks. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Nga, Cao Hong, Nguyen Khai Thinh, Pao-Chi Chang, and Jia-Ching Wang. Deep Learning Based Vietnamese Diacritics Restoration. In *2019 IEEE International Symposium on Multimedia (ISM)*, pages 331–3313. IEEE, 2019. doi: 10.1109/ISM46123.2019.00074.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, May 2020. European Language Resources Association. URL https://www.aclweb.org/anthology/2020.lrec-1.497.

Nuțu, Maria, Beáta Lőrincz, and Adriana Stan. Deep learning for automatic diacritics restoration in romanian. In *2019 IEEE 15th International Conference on Intelligent Computer Communication and Processing (ICCP)*, pages 235–240. IEEE, 2019. doi: 10.1109/ICCP48234.2019. 8959557.

Orife, Iroro. Attentive Sequence-to-Sequence Learning for Diacritic Restoration of YorùBá Language Text. *Proc. Interspeech 2018*, pages 2848–2852, 2018. doi: 10.21437/Interspeech. 2018-42.

Rychalska, Barbara, Dominika Basaj, Alicja Gosiewska, and Przemysław Biecek. Models in the Wild: On Corruption Robustness of Neural NLP Systems. In *International Conference on Neural Information Processing*, pages 235–247. Springer, 2019. doi: 10.1007/978-3-030-36718-3_ 20.

Straka, Milan, Jana Straková, and Jan Hajič. Czech Text Processing with Contextual Embeddings: POS Tagging, Lemmatization, Parsing and NER. In Ekštein, Kamil, editor, *Text, Speech, and Dialogue*, pages 137–150, Cham, 2019. Springer International Publishing. doi: 10.1007/978-3-030-27947-9_12.

Straková, Jana, Milan Straka, and Jan Hajič. Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-5003. URL `http://www.aclweb.org/anthology/P/P14/P14-5003.pdf`.

Zeman, Dan. DIAKRITIZACE TEXTU. In Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), editor, *CzechEncy - Nový encyklopedický slovník češtiny*. Nakladatelství Lidové noviny, Praha, Czech Republic, 2016.

Address for correspondence:
Jakub Náplava
`naplava@ufal.mff.cuni.cz`
Malostranské náměstí 25
118 00 Praha
Czech Republic