Diacritics Restoration for Slovak Texts Using Deep Neural Networks

Marek Šuppa[1]
Supervisor: prof. Ing. Igor Farkaš, Dr.[2]

[1] Katedra informatiky, FMFI UK, Mlynská Dolina, 842 48 Bratislava
[2] Katedra aplikovanej informatiky, FMFI UK, Mlynská Dolina, 842 48 Bratislava

June 8, 2018

Diacritics restoration – problem statement

Transform

> *Registraciu prace a odovzdanie suborov s prezentaciou spravite cez web.*

Diacritics restoration – problem statement

Transform

> *Registraciu prace a odovzdanie suborov s prezentaciou spravite cez web.*

to

> *Registráciu práce a odovzdanie súborov s prezentáciou spravíte cez web.*

Diacritics restoration – problem statement

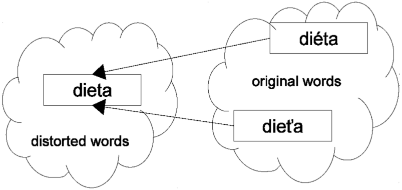
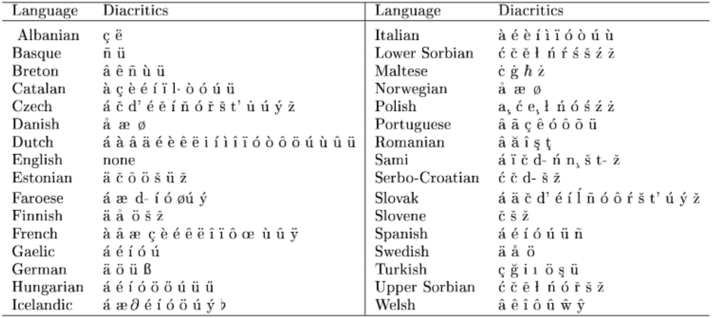


Figure: Ambiguous Mapping from Original to Distorted Set of Words [Hládek et al., 2013]



Language	Diacritics	Language	Diacritics
Albanian	ç ë	Italian	à é è í ì ï ó ò ú ù
Basque	ñ ü	Lower Sorbian	ć č ě ł ń ŕ ś š ź ž
Breton	â ê ñ ù ü	Maltese	ċ ġ ħ ż
Catalan	à ç è é í ï l· ò ó ú ü	Norwegian	å æ ø
Czech	á č d' é ě í ň ó ř š t' u̇ ú ý ž	Polish	a̧ ć ȩ ł ń ó ś ź ż
Danish	å æ ø	Portuguese	â ã ç ê ó ô õ ü
Dutch	á à â ä é è ê ë i í ì î ï ó ò ô ö ú ù û ü	Romanian	â ă î ş ţ
English	none	Sami	á ï č d- ń ņ š t- ž
Estonian	ä č õ ö š ü ž	Serbo-Croatian	ć č d- š ž
Faroese	á æ d- í ó øú ý	Slovak	á ä č d' é í Ĺ ñ ó ô ŕ š t' ú ý ž
Finnish	ä å ö š ž	Slovene	č š ž
French	à â æ ç è é ê ë î ï ô œ ù û ÿ	Spanish	á é í ó ú ü ñ
Gaelic	á é í ó ú	Swedish	ä å ö
German	ä ö ü ß	Turkish	ç ğ i ı ö ş ü
Hungarian	á é í ó ö ő ú ü ű	Upper Sorbian	ć č ě ł ń ó ř š ž
Icelandic	á æ∂ é í ó ö ú ý ♭	Welsh	â ê î ô û ŵ ŷ

Figure: Diacritics in European languages with Latin based alphabets. [Mihalcea and Nastase, 2002]

Related work

Diacritics restoration for Slovak texts:

- diakritik.korpus.sk [2014]
- Unsupervised spelling correction for Slovak [Hladek et al., 2013]
- Diacritics Restoration in the Slovak Texts Using Hidden Markov Model [Hládek et al., 2013] (appeared online 30 July 2016)

Related work

Diacritics restoration for Slovak texts:

- `diakritik.korpus.sk` [2014]
- Unsupervised spelling correction for Slovak [Hladek et al., 2013]
- Diacritics Restoration in the Slovak Texts Using Hidden Markov Model [Hládek et al., 2013] (appeared online 30 July 2016)

Related languages:

- Korektor for Czech [Richter et al., 2012]
- CzAccent–Simple Tool for Restoring Accents in Czech Texts [Rychlỳ, 2012]
- Corpus-Based Diacritic Restoration for South Slavic Languages [Ljubešic et al., 2016]
- Automatic diacritics restoration for hungarian [Novák and Siklósi, 2015]

Related work II

Diacritics restoration in general:

- Letter level learning for language independent diacritics restoration [Mihalcea and Nastase, 2002]
- Automatic diacritic restoration for resource-scarce languages [De Pauw et al., 2007]
- A word-based approach for diacritic restoration in Māori [Cocks and Keegan, 2011]
- Arabic Diacritization with Recurrent Neural Networks [Belinkov and Glass, 2015]

Is Diacritics Restoration an easy problem?

Number of all words	1 208 949
Number of unique words	899 702
Number of all 'clean' words	856 286
Words without alternations	515 245
LexDif score	1.05

Table: A sample of the statistics report for a subset of Slovak Wikipedia dump

Is Diacritics Restoration an easy problem?

Number of all words	1 208 949
Number of unique words	899 702
Number of all 'clean' words	856 286
Words without alternations	515 245
LexDif score	1.05

Table: A sample of the statistics report for a subset of Slovak Wikipedia dump

The number of word alternations is critical for diacritics restoration. For example, consider the latinized word `stat`

Is Diacritics Restoration an easy problem?

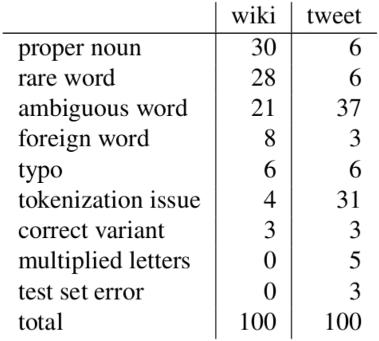
Number of all words	1 208 949
Number of unique words	899 702
Number of all 'clean' words	856 286
Words without alternations	515 245
LexDif score	1.05

Table: A sample of the statistics report for a subset of Slovak Wikipedia dump

The number of word alternations is critical for diacritics restoration. For example, consider the latinized word `stat`

- *stať* (a section)
- *štát* (a state)
- *sťať* (to cut down)
- *stáť* (to stand)

Ambiguity is usually the biggest problem



	wiki	tweet
proper noun	30	6
rare word	28	6
ambiguous word	21	37
foreign word	8	3
typo	6	6
tokenization issue	4	31
correct variant	3	3
multiplied letters	0	5
test set error	0	3
total	100	100

Figure: Error analysis on Slovene as provided in [Ljubešic et al., 2016]

General solution categorization

Published solutions seem to be categorizable depending on whether they operate on:

- word level
- sub-word level

Word level approaches

- Most often used approach
- Relies on big data corpuses
- Can further utilize other morphological tools

Word level approaches

- Most often used approach
- Relies on big data corpuses
- Can further utilize other morphological tools

A few examples:

- `diakritik.korpus.sk`
- Diacritics Restoration in the Slovak Texts Using Hidden Markov Model [Hládek et al., 2013]
- Corpus-Based Diacritic Restoration for South Slavic Languages [Ljubešic et al., 2016]

Word level approaches II

Sadly, these approaches quickly run into Zipf's law

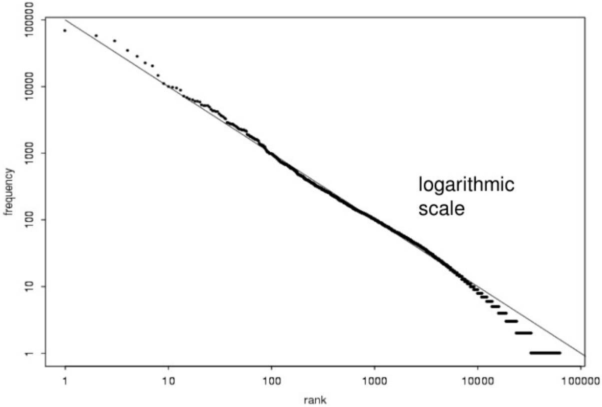


Figure: Illustration of Zipf's law on Brown Corpus (from [Manning et al., 1999])

Sub-word level approaches

- The usual approach for resource-scarce solutions
- Typically uses graphemes or their n-grams

Sub-word level approaches

- The usual approach for resource-scarce solutions
- Typically uses graphemes or their n-grams

A few examples:

- Automatic diacritic restoration for resource-scarce languages [De Pauw et al., 2007]
- Letter level learning for language independent diacritics restoration [Mihalcea and Nastase, 2002]
- Arabic Diacritization with Recurrent Neural Networks [Belinkov and Glass, 2015]

Deep Neural Networks for Diacritics Restoration

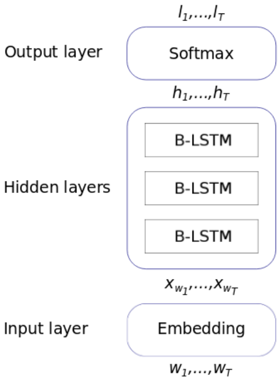


Figure: The network topology presented in [Belinkov and Glass, 2015]

Data

Data

- Slovak Wikipedia dataset
  35 613 022 words, 1 194 781 unique
- Digital Corpus of the European Parliament
  42 536 235 words, 713 273 unique

Results: Encoder type

encoder type	precision	recall	F1 score
RNN	0.859	0.917	0.885
LSTM	0.856	0.919	0.884
GRU	0.861	0.918	0.886
IndRNN	**0.944**	**0.947**	**0.941**

Table: A listing of results of evaluation of encoder types.

Results: Decoder type

decoder type	precision	recall	F1 score
''central only''	0.942	0.945	0.938
''flatten''	0.976	0.976	0.976
''attention''	**0.978**	**0.977**	**0.977**

Table: A listing of results of evaluation of decoder types.

Results: Best model

model	dataset	precision	recall	F1 score
baseline	Wikipedia	0.853	0.917	0.881
baseline	DCEP	0.851	0.917	0.881
our best model	Wikipedia	**0.987**	**0.989**	**0.988**
our best model	DCEP	**0.985**	**0.986**	**0.986**

Table: A listing of results of evaluation of our best model as compared to the baseline.

Confusion matrix

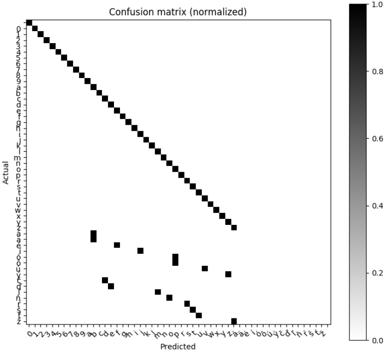


Figure: Confusion matrix of the baseline approach

Confusion matrix II

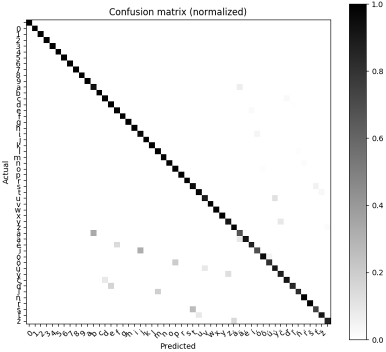


Figure: Confusion matrix of a trained model

Error analysis

input	metropolitnej oblasti gu$\boxed{\texttt{s}}$ dan v izraeli v relativ
predicted distribution	s: 0.8891944, š: 0.110016435, o: 0.00023678609
true output	š

Table: An example of a mistake made by our model which would require knowledge about a geographical location.

Error analysis II

input	iniciativ sucasnej etiky cnosti konkretne jej aris
predicted distribution	č: 0.6697172, c: 0.33018324, ŕ: 1.5026837e-05
true output	c

Table: An example of a mistake which is not straightforward even for Slovak native speakers.

Error analysis II

input	iniciativ sucasnej etiky $\boxed{c}$nosti konkretne jej aris
predicted distribution	č: 0.6697172, c: 0.33018324, ŕ: 1.5026837e-05
true output	c

Table: An example of a mistake which is not straightforward even for Slovak native speakers.

Experts from the Ľ. Štúr Institute of Linguistics of Slovak Academy of Sciences discuss this issue at https://jazykovaporadna.sme.sk/q/291/

Conclusions

- We analyzed previously published approaches to diacritics restoration.

Conclusions

- We analyzed previously published approaches to diacritics restoration.
- Based on this analysis, we designed new models based on Deep Learning and evaluated them on two datasets.

Conclusions

- We analyzed previously published approaches to diacritics restoration.
- Based on this analysis, we designed new models based on Deep Learning and evaluated them on two datasets.
- Our best model managed to achieve an 88% improvement over the baseline.

Future work

- Prepare an easy to use Open Source diacritics restoration tool based on the introduced model.

Future work

- Prepare an easy to use Open Source diacritics restoration tool based on the introduced model.
- Evaluate various other Deep Learning based models based on different approaches to input encoding.

Future work

- Prepare an easy to use Open Source diacritics restoration tool based on the introduced model.
- Evaluate various other Deep Learning based models based on different approaches to input encoding.
- Extend the scope of experiments to other languages.

Thank you for your attention!

Ba, J., Hinton, G. E., Mnih, V., Leibo, J. Z., and Ionescu, C. (2016).
Using fast weights to attend to the recent past.
In *Advances In Neural Information Processing Systems*, pages 4331–4339.

Belinkov, Y. and Glass, J. (2015).
Arabic diacritization with recurrent neural networks.
In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2281–2285.

Bradbury, J., Merity, S., Xiong, C., and Socher, R. (2016).
Quasi-recurrent neural networks.
arXiv preprint arXiv:1611.01576.

Cocks, J. and Keegan, T. T. (2011).
A word-based approach for diacritic restoration in māori.
In *Australasian Language Technology Association Workshop 2011*, page 126.

- De Pauw, G., Wagacha, P. W., and De Schryver, G.-M. (2007).
  Automatic diacritic restoration for resource-scarce languages.
  In *Text, Speech and Dialogue*, pages 170–179. Springer.

- Hládek, D., Staš, J., and Juhár, J. (2013).
  Diacritics restoration in the slovak texts using hidden markov model.
  In *Language and Technology Conference*, pages 29–40. Springer.

- Hladek, D., Stas, J., and Juhar, J. (2013).
  Unsupervised spelling correction for slovak.
  Advances in Electrical and Electronic Engineering, 11(5):392.

- Hládek, D., Staš, J., and Juhár, J. (2014).
  Slovak web discussion corpus.
  In *International Conference on Natural Language Processing*, pages 463–469. Springer.

- Ljubešic, N., Erjavec, T., and Fišer, D. (2016).
  Corpus-based diacritic restoration for south slavic languages.
  In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA)(may 2016).*

- Luong, M.-T., Le, Q. V., Sutskever, I., Vinyals, O., and Kaiser, L. (2015).
  Multi-task sequence to sequence learning.
  arXiv preprint arXiv:1511.06114.

- Manning, C. D., Schütze, H., et al. (1999).
  Foundations of statistical natural language processing, volume 999.
  MIT Press.

- Mihalcea, R. and Nastase, V. (2002).
  Letter level learning for language independent diacritics restoration.

In *proceedings of the 6th conference on Natural language learning-Volume 20*, pages 1–7. Association for Computational Linguistics.

Novák, A. and Siklósi, B. (2015).
Automatic diacritics restoration for hungarian.
Association for Computational Linguistics.

Richter, M., Straňák, P., and Rosen, A. (2012).
Korektor–a system for contextual spell-checking and diacritics completion.
In Kay, M. and Boitet, C., editors, *Proceedings of the 24th International Conference on Computational Linguistics (Coling 2012)*, pages 1–12, Mumbai, India. IIT Bombay, Coling 2012 Organizing Committee.

Rychlỳ, P. (2012).
Czaccent–simple tool for restoring accents in czech texts.

RASLAN 2012 Recent Advances in Slavonic Natural Language Processing, page 85.

Re: Visualization of Embeddings

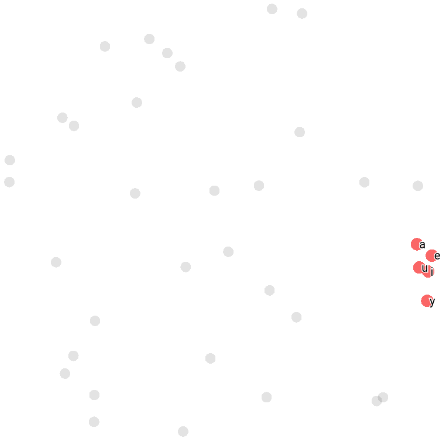


Figure: A visualization of the embedding layer learned as part of training the Diacritics Restoration model.

Re: Upper bound for restoration accuracy

input	`iniciativ sucasnej etiky` $\boxed{\texttt{c}}$`nosti konkretne jej aris`
predicted distribution	č: 0.6697172, c: 0.33018324, ŕ: 1.5026837e-05
true output	c

Table: An example of a mistake which is not straightforward even for Slovak native speakers.

Strangely enough, `diakritik.korpus.sk` correctly predicts ”c” in this case, which is most probably due to the fact that ”cnosť” is found more often in the texts that comprised this model's training set.

Re: IndRNN's effectiveness

No simple answer yet. A very new model (introduced in March 2018), a lot more exploration is needed.