CoLi – Final Project – Diacritics Restoration

Herbert Ullrich (**2576412**)

March 25, 2019

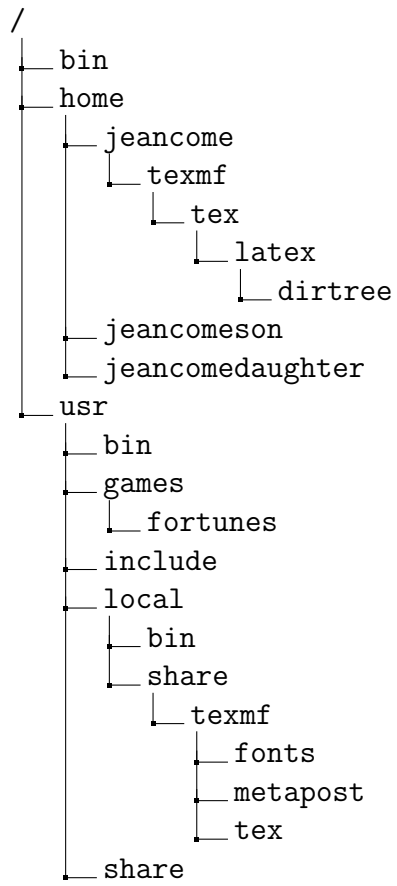
1 Introduction

Diacritic restoration is the task to restore diacritics, novel algorithms based on recurrent neural networks such as [2]

1.1 Used corpora

We have decided to use the *Corpus for training and evaluating diacritics restoration systems* [1] proposed in 2018 as the standard training and testing set for the DR task. That way, our results will get easy to compare with the state-of-the-art methods, such as [2]

1.2 Project structure



language	train. sentences	dia-words	n	acc.	word acc.	dia-word acc.
Croatian	802,610	14.3%	1	0.977	0.857	0
			2	0.98	0.874	0.206
			3	0.985	0.903	0.491
			4	**0.989**	**0.929**	**0.618**
Czech	952,909	47.8%	1	0.892	0.522	0
			2	0.904	0.549	0.141
			3	0.924	0.628	0.321
			4	**0.94**	**0.689**	**0.433**
Slovak	613,727	41.1%	1	0.922	0.589	0
			2	0.925	0.614	0.123
			3	0.938	0.669	0.274
			4	**0.951**	**0.727**	**0.42**
Irish	50,825	29.2%	1	0.943	0.708	0
			2	0.948	0.736	0.214
			3	0.957	0.782	0.407
			4	**0.962**	**0.802**	**0.445**
Hungarian	1,294,605	47.4%	1	0.906	0.526	0
			2	0.906	0.52	0.0792
			3	0.924	0.587	0.267
			4	**0.939**	**0.649**	**0.38**
Polish	1,069,841	31.6%	1	0.946	0.684	0
			2	0.949	0.701	0.164
			3	0.961	0.762	0.353
			4	**0.969**	**0.809**	**0.493**
Romanian	837,647	28.2%	1	0.95	0.723	0.0234
			2	0.953	0.741	0.164
			3	0.958	0.768	0.31
			4	**0.964**	**0.797**	**0.431**
French	1,818,618	16.6%	1	0.97	0.834	0
			3	0.971	0.84	0.0933
			4	**0.975**	**0.861**	**0.274**
Spanish	1,735,516	11.8%	1	0.981	0.882	0
			3	0.982	0.893	0.218
			4	**0.985**	**0.909**	**0.378**
Latvian	315,807	46.8%	1	0.915	0.532	0
			3	0.924	0.589	0.24
			4	**0.939**	**0.654**	**0.38**

Table 1: Measurements of per-tag (single diacritical mark) and per-word accuracy of an n-gram based HMM diacritic restorer. We call dia-word a word that contains at least one diacritical character. The dia-word percentage is listed for testing set.

References

[1] Jakub Náplava, Milan Straka, Jan Hajič, and Pavel Straňák. Corpus for training and evaluating diacritics restoration systems, 2018. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[2] Jakub Náplava, Milan Straka, Jan Hajič, and Pavel Straňák. Diacritics Restoration Using Neural Networks. In Nicoletta Calzolari (Conference chair), Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA).