A Appendices

A.1 Hyperparameters

Table 6 reports the best obtained hyperparameters for models trained on `text8` corpus. These are minimum count (MC), negative sample rate (NS), epochs (EP), learning rate (lr), and optimiser (Opt.). For models trained on WN18RR hyperparameter where identical to the ones indicated in the original works, as ide from negative samples (best obtain 10) and epochs, kept at 50, as indicated in the paper. Results Obtained on the WN18RR test split did not significantly differ form the scores reported in the original works. Again, the total set of parameters was obtain by intersecting the ones presented in the models' original papers (Czarnowska et al., 2019; Balazevic et al., 2019; Chami et al., 2020).

	MC	NS	EP	lr	Opt.
DM	100	20	5	.001	Adam
MuRE	0	40	50	50	SGD
RotE	0	30	15	50	SGD
RefE	0	30	15	50	SGD
AttE	0	25	10	50	SGD

Table 6: Best hyperparameters for models trained on `text8` corpus.

A.2 Vocabulary Coverage

We here present the final coverage for all the benchmarks used for the models trained on the WN18RR (Table 8) and `text8` (Table 7) corpora.

Benchmark	Coverage
SimLex	726/999
MEN	1544/3000
WS353_sim	152/203
WS353_rel	200/251
ML10 Adjective Nouns	1836/1944
ML10 Verb Objects	1836/1944
ML10 Noun-Nouns	1782/1944

Table 7: Final coverage of the different datasts' items used for testing models trained on `text8`.

Note the significantly smaller coverage that models trained on WN18RR show for Adjective Noun phrases on Table 8. Such small coverage is one of the main reason that guided the decision towards not sharing the word vocabulary across models trained on the two different corpora.

Benchmark	Coverage
SimLex	787/999
MEN	1635/3000
WS353_sim	166/203
WS353_rel	200/251
ML10 Adjective Nouns	648/1944
ML10 Verb Objects	1674/1944
ML10 Noun-Nouns	1494/1944

Table 8: Final coverage of the different datasts' items used for testing models trained on WN18RR.

A.3 Statistical Significance

We here report those Model-Strategy pairs for which the observed differences in the correlation analysis are not statistically significant, according to our bootstrap test.

Phrase Type	**Model A**	**Model B**	***p***
NN	DM-add	DM-Rt	.728
NN	DM-Rh	DM-Rt	.216
VO	DM-add	DM-Rh	.864
NN	DM-add	DM-BiD	.066
NN	DM-add	DM-Rh	.213
NN	DM-add	DM-Rt	.410
NN	DM-Rh	DM-Rt	.268
VO	DM-add	DM-Rt	.147

Table 9: Bootstrap analyses results, stratified by different random seeds. *p* values refers to Holm-corrected values.

A.4 Single Space DM

We are aware that Zobnin and Elistratova (2019) proposed a method to reduce SGNS vector spaces to one, and run a few preliminary experiments adopting this strategy in DM. As presented in Figure 3, such experiments clearly suggest that DM is superior to the investigated variants.

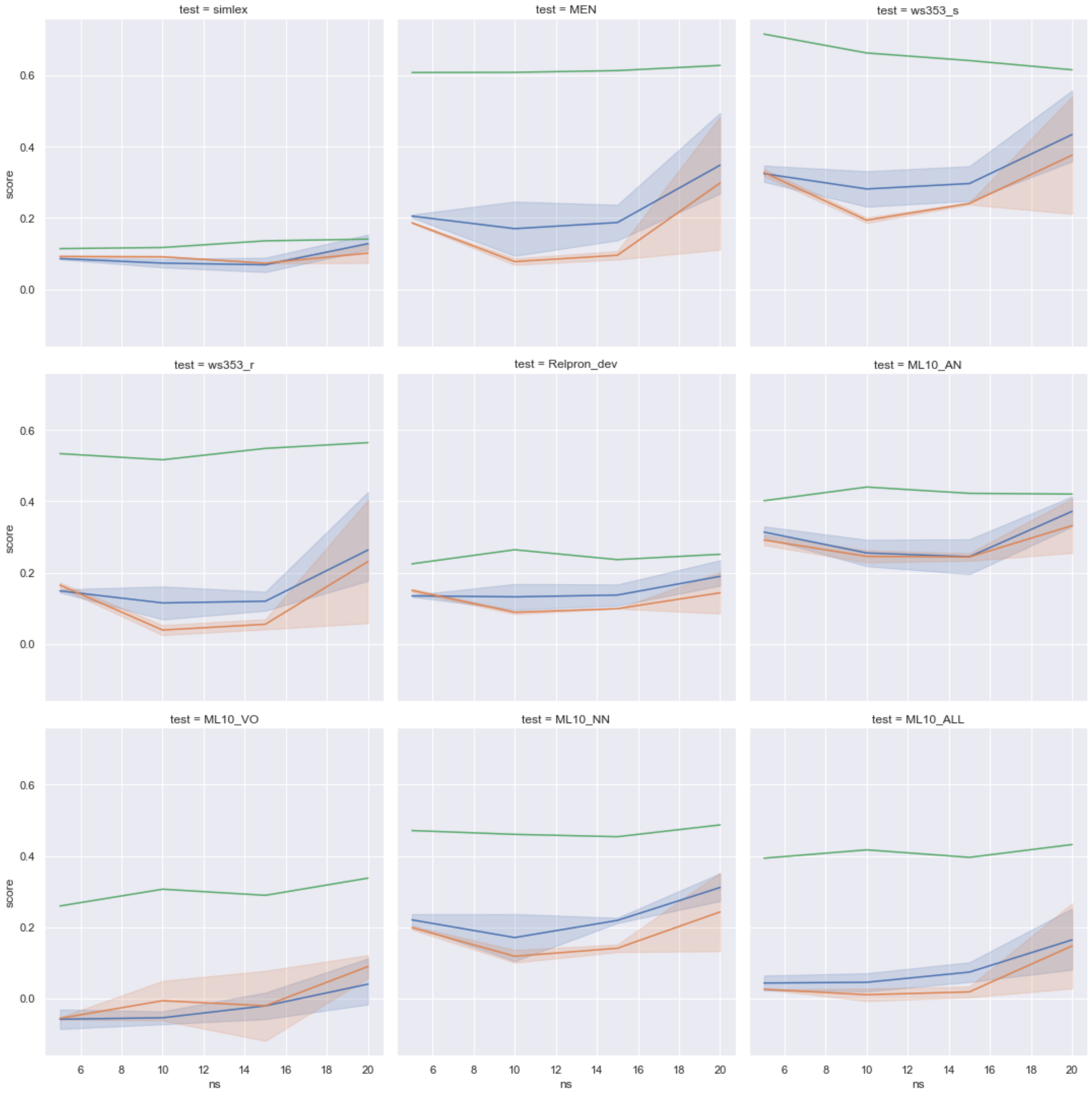

Figure 3: Comparison of results on all the benchmarks discussed in the paper with a DM model and two single-space version, OSDM and FullOSDM, obtained applying Zobnin and Elistratova (2019) method to the DM. The shaded areas refer to the fact that these models included the extra hyperparameter q.