e2e-Dutch model trained on Sonar-1

This is a model to be used with the e2e-Ducth code, trained on the SoNaR corpus.

Training dataset

The data for training this model was taken from the SoNaR-1 corpus [1], which is a dataset of 1 million words of contemporary written Dutch. The text is annotated with several linguistic features, among which coreference resolution. For this model, only the coreference resolution annotations were used.

The tool mmax2conll was used to convert the SONAR MMAX files to CONLL-2012 style files. The dataset was split into a train, validation and test set (individual documents were not split). Train, test and validation sets were each joined together into one large CONLL-2012 file.

Embedding files

For word embeddings, a fasttext model of dimension 300 was used [2]. This model can be downloaded here was used.

For the contextual embeddings, Bertje [3] is used. This model is available directly from the `transformers` library.

Other settings

The precise model settings can be found in the attached config file.

References

[1]: SoNaR-corpus (Version 1.2.1) (2015) (Data set). Available at the Dutch Language Institute: http://hdl.handle.net/10032/tm-a2-h5

[2]: Grave, Edouard, et al. "Learning word vectors for 157 languages." *arXiv preprint arXiv:1802.06893* (2018).

[3]: de Vries, Wietse, et al. "Bertje: A dutch bert model." *arXiv preprint arXiv:1912.09582* (2019).