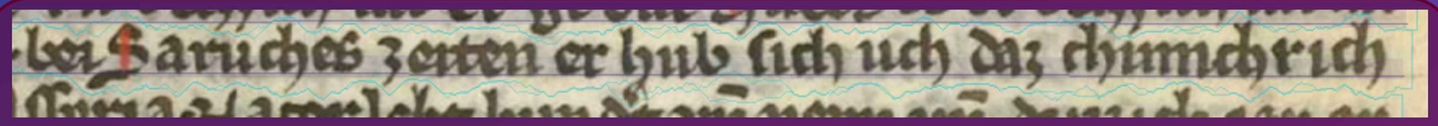
Hydra.

(universal) Tagger -Lemmatiser with Deep Learning and Parallel Computing

Premodern spellling

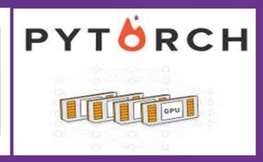
In the field of historical computer linguistics we invariably face a difficulty related to reliable (pre-)processing, more precisely speaking, tagging and lemmatisation of premodern texts. The problem is the nature of premodern spelling, both in Latin and in the vernacular languages of Europe, exposing a high degree of variance (between regions and even between particular scribes, as far as scribal culture is considered) that inevitably impedes the automatic text processing on scale. Not only a single word may have been spelled differently, but the boundaries between words and morphemes were changing from one manuscript to another. That fluidity in orthographic practice probably forms the biggest impediment to natural language processing of premodern historical texts.

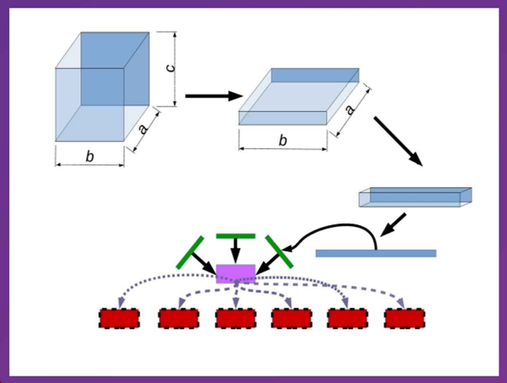


Multi-token single label problem. In this Middle High German sentence *bei Saruches zeiten er hub sich uch das chunichrich [in Assyria]* ['in the time of Serug there arose also the kingdom [of Assyria]'] the reflective verb *er hub sich* [arose, emerged] is presented by two tokens in the diplomatic transcription widely used in medieval studies.

Multi-label single token problem. In this Middle Low German sentence *dat hauestu ghedan* ['you did it/have done it'] the token *hauestu* is actually the enclitic form of *hauest tu* and needs two labels (part-of-speech tags and lemmas for both the verb *hauest* '(you) have' and the personal pronoun *tu* 'you').

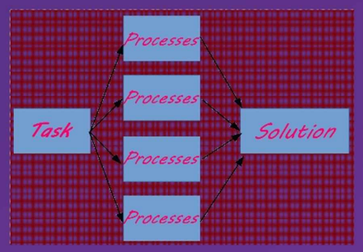
Deep Learning



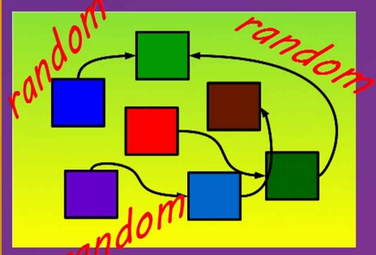


Multi-layer Artificial Neural Networks aka Deep Learning is a field of research aiming at complex learning methods that can efficiently analyse huge amount of data and solve complicated non-trivial tasks. By representing words and whole phrases or sentences as embedding tensors we compress and mine relevant linguistic information concealed, for instance, in affixes or words combinations. In this way we ensure high accuracy of a label classification depending both on the word structure and the word context.

Parallel Computing and Gossip Algorithm

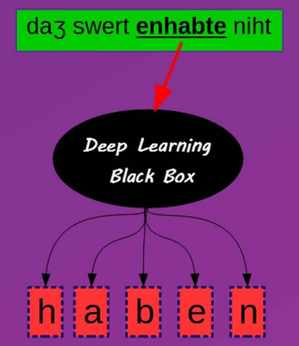


Parallel computing is an umbrella term for different methods of sharing data and computations among many separate units (processors, graphic cards, machines etc.). One of the strategies for the coordination of parallel processes is the so-called gossip algorithm which allows processes to exchange relevant information between each other in a random way. The final solution is an average of all the processes' solutions that are supposed to have converged in the meantime.



Lemma Building

Since large corpora may encompass tens of thousands of lemmas, it can be also quite difficult to compute model probabilities for all of them, especially for those less frequent. An experimental approach to construct a lemma from scratch is computationally easier and is thought to predict lemmas absent in the training corpus by looking at analogies already observed.



Selected Literature

- Blot M., Picard D., Cord M., Thome N. 'Gossip training for deep learning'. NIPS 2016 workshop, Barcelona, Spain, December 2016. https://arxiv.org/abs/1611.09726
- Kestemont, M., De Pauw, G., Van Nie, R. & Daelemans, W., 'Lemmatisation for Variation-Rich Languages Using Deep Learning'. *DSH – Digital Scholarship in the Humanities* 32:4 (2017), 797-815. DOI: https://doi.org/10.1093/llc/fqw034
- Kestemont, M. & J. de Gussem, 'Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning', *Journal of Data Mining & Digital Humanities* (2017), pp. 17. Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages, ed. M. Buechler and L. Mellerin.
- Piotrowski M., *Natural Language Processing for Historical Texts*, San Rafael: Morgan & Claypool Publishers, 2012
- Zhang Y., Shen D., Wang G., Gan Z., Henao R., Carin L. 'Deconvolutional Paragraph Representation Learning', Conference on Neural Information Processing Systems 2017, Long Beach. arXiv:1708.04729

https://github.com/Lukasz-G/Hydra

gefördert durch Cusanuswerk Bischöfliche Studienförderung

Georg-August-Universität Göttingen