Visualizing Language Models

Tobias Hodel, Sebastian Flick, Reinhard Priber, Christa Schneider, Jonas Widmer
Digital Humanities, dh@wbkolleg.unibe.ch

Introduction

Language models are essential to succeed in NLP tasks such as Part-of-Speech-Tagging and Named Entity Recognition. Since the advent of *word2vec* and large transformer-based language models (such as *BERT* or *GPT-3*) a variety of specialized and fine-tuned language models is currently available. Nonetheless, our in-depth understanding of such models is limited at best.

In order to strengthen our knowledge of language models and to start the process of reflecting them, we are searching for creative ways of visualizing language models. We envision 3D-visualizations based on dimension reduction to identify the positioning of homonyms in vector spaces.

In our understanding, homonyms are words which share the same spelling but not the same meaning (homographs). Until today, this kind of homonyms is known to cause difficulties when processed in language models, as they are treated as one single element regardless to their semantic differences.

By using current frameworks [Akbik et al. 2019] which embed tokens ("words") as parts of sentences, semantic and grammatical context plays an important role. Based on these contexts, homographs can for example be compared, when visualized even physically.

This leads us to new discussions and first very simple explanations of the inner functions of language models. As such models are currently implemented in tools (like search engines) and machines (like smartphones) that we use daily, we can expect to deal with other, even more complex questions.

Methods/Workflow

In order to make language models visually approachable we:

- Based on textual data from the Königsfelden corpus (for details see below)
- Trained a forward and backward character-based language model (vectorization of textual data) from scratch using *flair NLP*, not based on any pre-trained models [2096 dimensions]
- Vectorized 30 sentences containing homographs from the corpus
- Stored in a database for further use
- Reduced to three dimensions by applying PCA
- Visualized in a svelte frontend
- Deployed as a docker-image

Code

URL: https://github.com/DHBern/lm_viz

Dataset

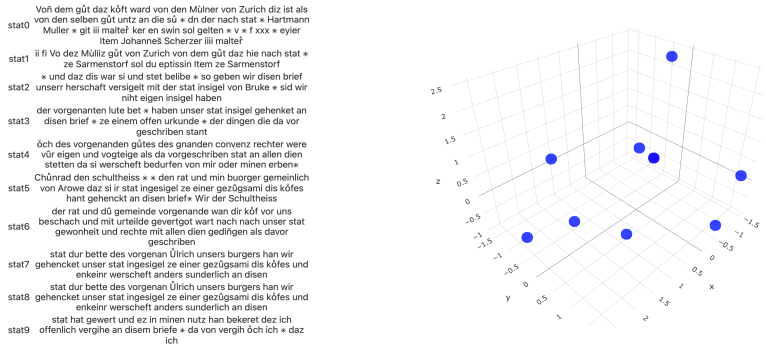
URL: https://koenigsfelden.uzh.ch

Charters and Records of Königsfelden Abbey (1308-1662). Prepared as a scholarly edition. Halter-Pernet et al. 2021: https://koenigsfelden.uzh.ch.

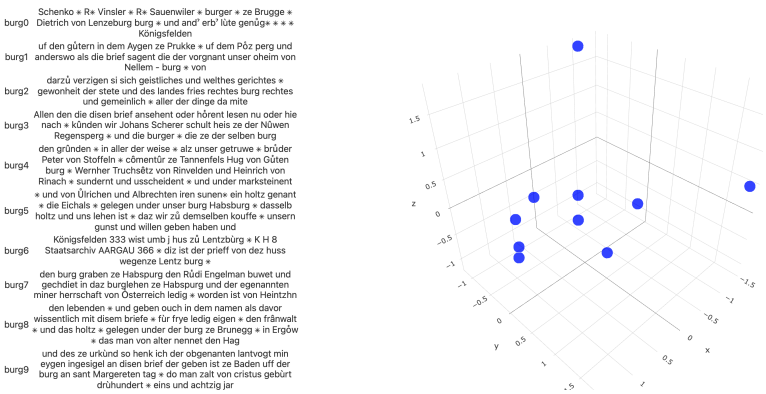
Results

URL: https://nlp-hack-4.fdn-dev.iwi.unibe.ch

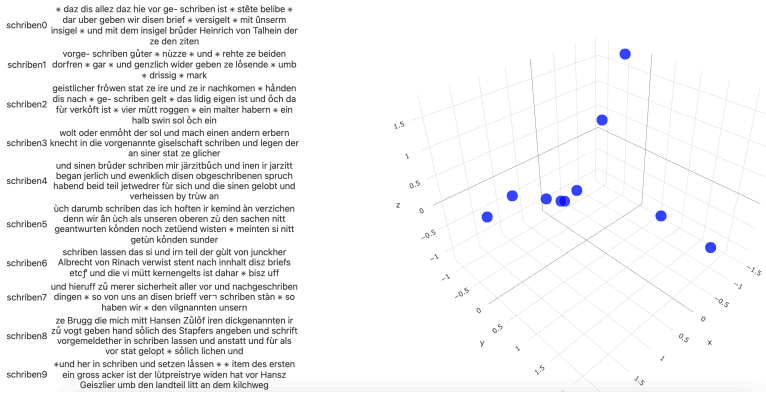
Label 1 "stat" (*city, place of execution, instead*)



Label 2 "burg" (*castle, person & place name*)



Label 3 "schriben" (*written, writing*)



Digital Humanities
Universität Bern
Walter Benjamin Kolleg
Mittelstrasse 43
3012 Bern