Abstract

Johan Jarlbrink & Roger Mähler, Umeå University

Embedded words in the historiography of technology and industry, 1931–2016

Short paper

From 1931 to 2016 The Swedish National Museum of Science and Technology published a yearbook, *Dædalus*. The 86 volumes display a great diversity of industrial heritage and cultures of technology. The first volumes were centered on the heavy industry, such as mining and paper plants located in North and Mid-Sweden. The last volumes were dedicated to technologies and products in people's everyday lives – lipsticks, microwave ovens, and skateboards. During the years *Dædalus* has covered topics reaching from individual inventors to world fairs, media technologies from print to computers, and agricultural developments from ancient farming tools to modern DNA analysis. The yearbook presents the history of industry, technology and science, but can also be read as a historiographical source reflecting shifting approaches to history over an 80-year period. *Dædalus* was recently digitized and can now be analyzed with the help of digital methods.

The aim of this paper is twofold: To explore the possibilities of word embedding models within a humanities framework, and to examine the *Dædalus* yearbook as a historiographical source with such a model. What we will present is work in progress with no definitive findings to show at the time of writing. Yet, we have a general idea of what we would like to accomplish. Analyzing the yearbook as a historiographical source means that we are interested in what kinds of histories it represents, its focus and bias. If words are defined by the distribution of the vocabulary of their contexts we can calculate relations between words and explore fields of related words as well as binary relations in order to analyze their meaning. Simple – and yet fundamental – questions can be asked: What is "technology" in the context of the yearbook? What is "industry"? Of special interest in the case of industrial and technological history are binaries such as rural/urban, man/woman, industry/handicraft, production/consumption, and nature/culture. Which words are close to "man", and which are close to "woman"? Which aspects of the history of technology and industry are related to "production" and which are related to "consumption"?

Word embedding is a comparatively new set of tools and techniques within data science (NLP) with that in common that the words in a vocabulary of a corpus (or several corpora) are assigned numerical representations through some (of a wide variety of different) computation. In most cases, this comes down to not only mapping the words to numerical vectors, but doing so in such a way that the numerical values in the vectors reflect the contextual similarities between words. The computations are based on the distributional hypothesis stemming from Zellig Harris (1954), implicating that "words which are similar in meaning occur in similar contexts" (Rubenstein & Goodenough, 1965). A predecessor of the method used here is the self-organizing maps, where the contextual roles of words are represented in a two-dimensional space. (Kohonen, 1995). With word embedding the words are embedded (positioned) in a high-dimensional space, each word represented by a vector in the space i.e. a simple representational model based on linear algebra. The dimension of the space is defined by the size of the vectors and the similarity between words then become a matter of computing the difference between vectors in this space, for instance the difference in (euclidian) distance or difference in direction between the vectors (cosine similarity). Within vector space models the former is the most popular under the assumption that related words tend to have similar directions. The arguably most prominent and popular of these algorithms, and the one that we have used, is the skip-gram model Word2Vec (Mikolov et al, 2013). In short, this model uses a neural network to compute the word vectors as results from training the network to predict the probabilities of all the words in a vocabulary being nearby (as defined by a window size) a certain word in focus.

An early evaluation shows that the model works fine. Standard calculations often used to evaluate the performance and accuracy indicates that we have implemented the model correctly – we can indeed get the correct answers to equations such as "Paris - France + Italy = Rome" (Mikolov et al, 2013). In our case we were looking for "most_similar(positive=['sverige','oslo'], negative=['stockholm'])". And the "most similar" was "norge". We have also explored simple word similarity in order to evaluate the model and get a better understanding of our corpus. What remains to be done is to identify relevant words (or group of words) that can be used when we are examining "topics" and binary dimensions in the corpus. We are also experimenting with different ways to cluster and visualize the data. Although some work remains to be done, we will definitely have results to present at the time of the conference.

Harris, Zellig (1954). Distributional structure. Word, 10(23):146–162.

Kohonen, Teuvo (1995). *Self-organizing maps*. Berlin: Springer

Mikolov, Tomas, Chen, Kai, Corrado, Greg & Dean, Jeffrey (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781

Rubenstein, Herbert & Goodenough, John (1965). Contextual Correlates of Synonymy. *Communications of the ACM*, 8(10): 627-633.