Natural language processing (NLP) for register-based research

Novel descriptions and improved causal inference

Martin Arvidsson & Benjamin Jarvis

Background & motivation

- Increased availability of large-scale digitized corpora has sparked a growing interest within the social sciences for natural language processing (NLP) methods. Two families of methods in particular have received special attention; *topic models* [1] and *word embedding models* [2], which have enabled social scientists to extract novel and interpretable patterns from textual data.

- The central idea behind both topic models and word embedding models is to infer *meaning of words* based on the *contexts* in which they occur.

- For **topic models**, context is typically defined at the *document*-level, and the basic idea is to estimate "topics" (distribution over words) such that words which frequently co-occur in documents are allocated to the same topic(s)[1].

- For **word embedding models**, context is typically defined in terms of the ±*k surrounding words*, and the basic idea is to estimate vector representations (embeddings) of each word such that embeddings of context words are predictive of the focal word. Hence, words occurring in similar contexts become allocated close in vector space, and vice versa.

- These methods have been shown to *encode surprisingly rich semantic structures*. For example, by subtracting the embedding of "man" from "king", and adding "woman", the closest word in embedding space often is "queen"; illustrating how both *gender*- and *royalty* dimensions are encoded into such embeddings.

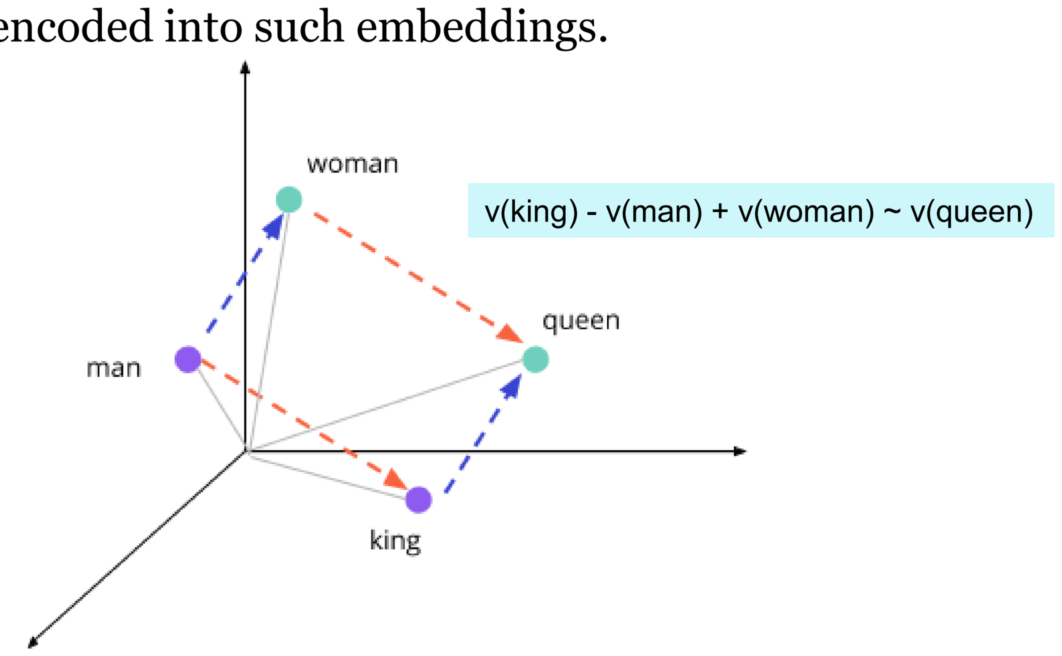

Figure 1 Toy example of embedding space [3]

Research question

Can the successful application of these methods be replicated for other kinds of high-dimensional data that interest social scientists?

- Can we, instead of inferring properties about *words*, infer properties about *individuals* from their *social contexts* (see Figure 2)?
- Can latent structures be encoded such that ---just as for textual data--- (a) novel descriptions and (b) improved causal inference can be achieved?

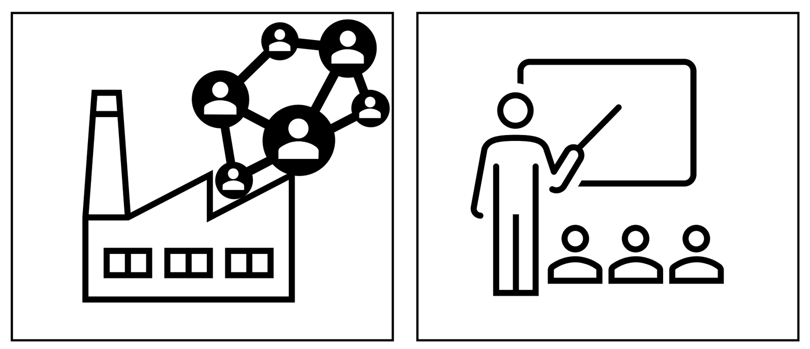

Figure 2 Examples of *social contexs*: workplaces & schools.

Some preliminary results

- In this poster presentation, I present two sets of preliminary results based on Swedish register data:

- (1) Conceptualizing *workplaces* as documents and *individuals'* educational status (*Sun2000Inr*) as words, we use topic models to investigate how *employment structures* have changed over time in Stockholm.

- (2) Conceptualizing *"neighboring workplaces"* in a *mobility network* as contexts, and *workplaces* as words, we use word embeddings to proxy unobserved drivers of mobility when estimating peer effects.

Results (1)

- Figure 3 shows that the estimated topics appear to have good face-validity; educational-labels that intuitively are related have been grouped together.

Topic 1	Topic 19	Topic 39
1. Byggnadsarbete	1. Illustration, reklam, grafisk formgivning & foto	1. Fartygs- och flygteknik
2. Byggnadssnickeri	2. Annan utbildning inom medieproduktion	2. Utbildning för sjöfart
3. Byggnads- och anläggningsteknik	3. Journalistik	3. Annan utbildning inom transporttjänster
4. Trätekniskt arbete	4. Journalistik & medievetenskap	4. Utbildning inom luftfart
5. Betong-, anläggning & vägarbete	5. Marknadsföring	5. Ingenjörsutbildning, inrikt. fordon & farkostteknik

Figure 3 Most likely education-labels in three topics.

- Figure 4 provides one example of how, by examining the temporal composition of topics, we can learn about how individuals' working environment changed over time; In the 90s, the most likely topic for *social workers* was one consisting of individuals educated in *psychiatric care*. In the beginning of the 2000s, however, *social workers* became the most defining educational-label of another topic, one consisting of individuals educated in *social science* and *administration*.

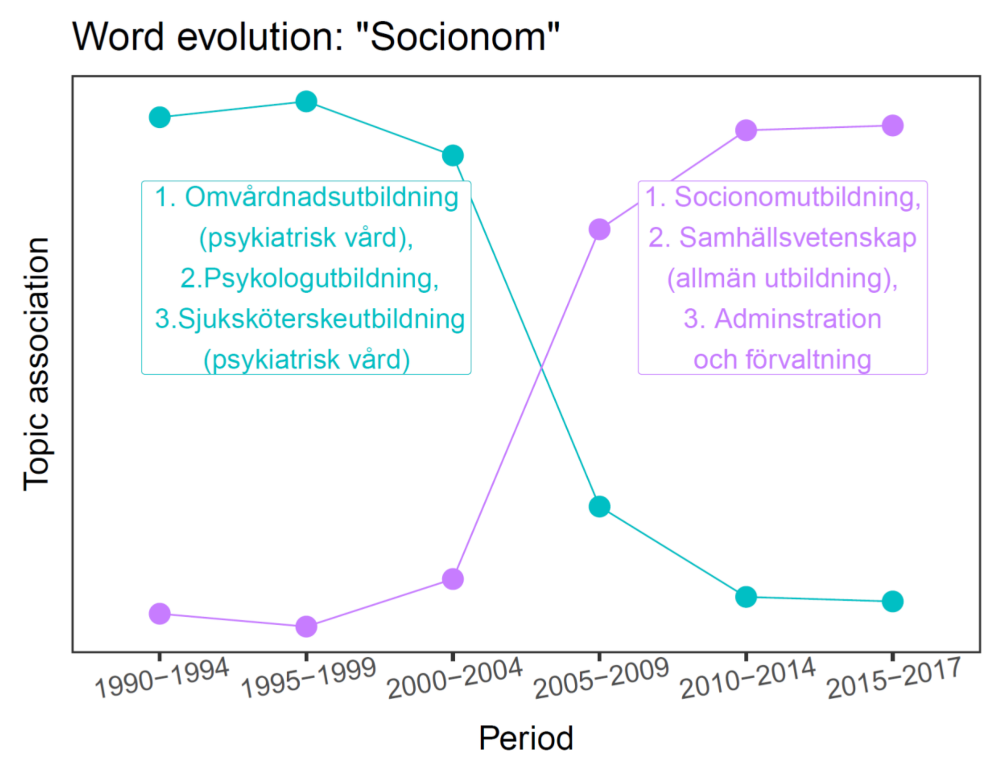

Figure 4 Example of change in topic association.

- Figure 5 shows the temporal prevalence of the topic with the biggest relative drop (Data/Finance ~2000) as well as it most closely related topic (Data/Engineering). What happened around ~2000? The dot-com bubble. This type of analysis ---examining how aggregated topic proportions change over time--- could be used more generally to study emerging/declining co-occurrence patterns (indicative of e.g., emerging industries).

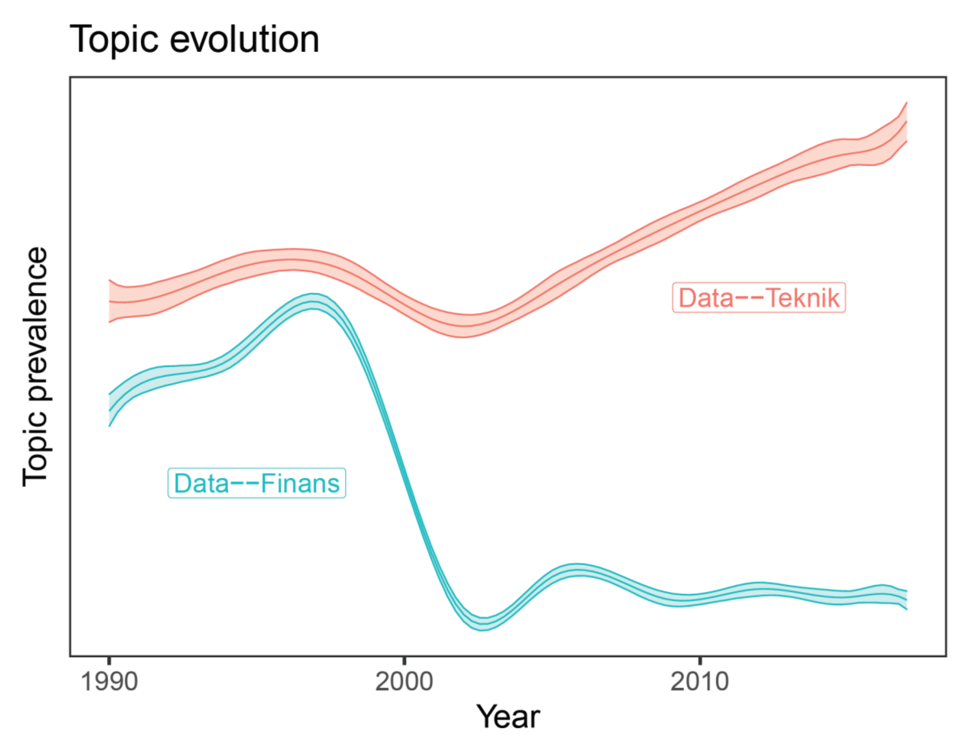

Figure 5 Prevalence of two "data"-related topics.

Results (2)

- Does the presence of a *prior move* between two firms <*j*,*k*> increase the probability of a *future move*? To answer this question, we use a matching design (Figure 6), and identify counterfactual pairs <*i*,*k*> where *i* is a firm very similar to *j* but which did not have a prior move to *k*.

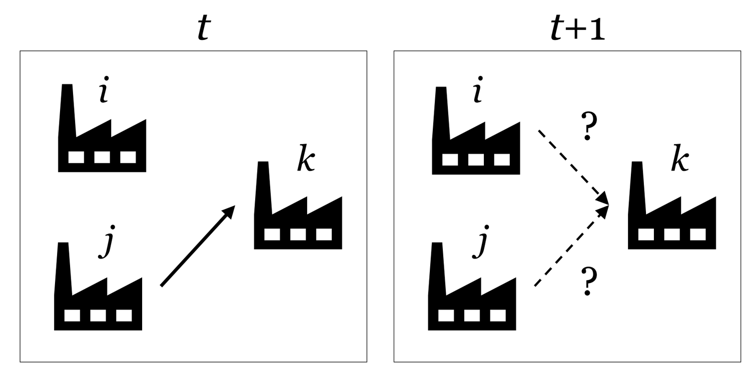

Figure 6 Matching design.

- Figure 7 shows that matching on *embeddings* reduces treatment effects substantially (by ~33%) compared to matching *only on observables*.

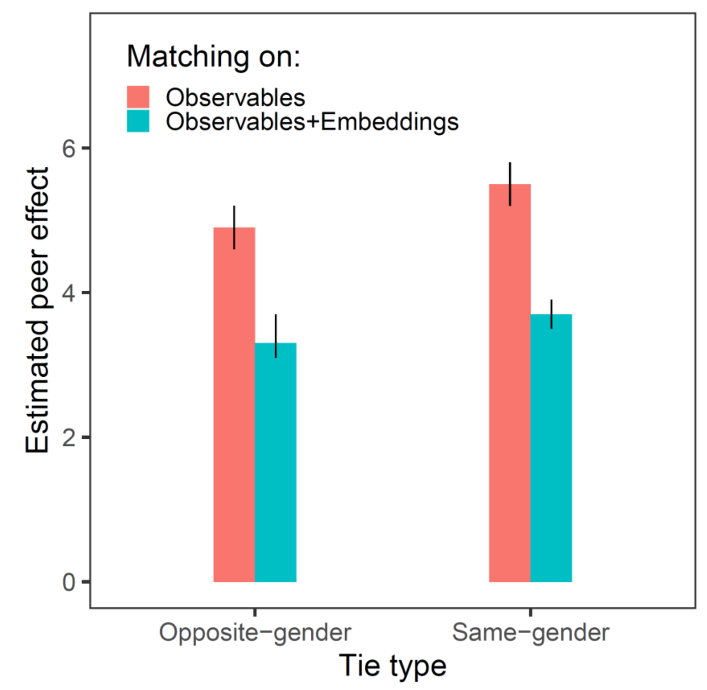

Figure 7 The effect of *prior* moves on *future* moves.

Conclusions

- These results suggest that NLP methods indeed can be useful for other types of social data and encode latent structures such that both novel descriptions & improved causal inference can be attained.

Footnotes

[1] It is important to underscore that topics are not predefined but *discovered*.

References

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.
2. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
3. https://bit.ly/2P8Kvcc

LINKÖPINGS UNIVERSITET

LINKÖPINGS UNIVERSITET
Institutet för Analytisk Sociologi