Blindfolded NLP: Unsupervised Learning for Automatically Generating Topic Labels

Charley Wu[1], Zsolt Jurányi[2], Laszlo Gulyas[2], George Kampis[2]

[1]Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development

[2]Petabyte Research Ltd, Budapest

Abstract

The use case of our study is the Hungarian Internet Archive pilot, which was used to create a distributional semantic model of the Hungarian language. Using unsupervised methods, documents from the corpus were grouped into semantically related clusters. Then topic labels were automatically generated for each cluster by fitting a probability distribution to each cluster. Query vectors were sampled from the probability distribution and used to search the semantic space of the language model to yield the terms with the highest semantic relevance. Results are assessed for the applicability of the method for automated semantic labeling and topic detection.