Visualisation and Exploration of High-Dimensional Distributional Features in Lexical Semantic Classification

Maximilian Köper¹, Melanie Zaiß², Qi Han², Steffen Koch², Sabine Schulte im Walde¹

¹Institut für Maschinelle Sprachverarbeitung
²Institut für Visualisierung und Interaktive Systeme
Universität Stuttgart, Germany

¹{maximilian.koeper,schulte}@ims.uni-stuttgart.de
²{firstname.lastname}@vis.uni-stuttgart.de

Abstract

Vector space models and distributional information are widely used in NLP. The models typically rely on complex, high-dimensional objects. We present an interactive visualisation tool to explore salient lexical-semantic features of high-dimensional word objects and word similarities. Most visualisation tools provide only one low-dimensional map of the underlying data, so they are not capable of retaining the local and the global structure. We overcome this limitation by providing an additional trust-view to obtain a more realistic picture of the actual object distances. Additional tool options include the reference to a gold standard classification, the reference to a cluster analysis as well as listing the most salient (common) features for a selected subset of the words.

System Overview

- **A** Main Window: The main window presents the visualisation of the T-SNE-reduced two-dimensional data points. The main window enables the user to get an overview of the spatial locations of the target objects, and their distances from each other.
- **B** Navigation Bar: allows saving and loading previous coordinates as well as generating a new visualisation (Refresh). Further one can select another underlying feature set and gold standard.
- **C** Goldclass Assignments: displays the gold standard class labels. Selecting a label marks all elements in window [A] with the same colour.
- **D** Optional Cluster Assignments: if a cluster analysis file is loaded, this window allows highlighting cluster memberships.
- **E** Common Features Display: When selecting an individual element in the main window [A], the entire main window freezes. Clicking on an additional element then displays common features of the selected elements in window [E], i.e. words that co-occur with both elements. These features are sorted according to feature scores.

Trust-View

- Dimensional reduction comes at the cost of information loss, leading to distortions in the displayed distances of vectors in the low dimensional space.
- To diminish this problem, the trust-view lets users inspect a vector interactively by hovering it with the mouse.
- In response to this interaction, all other word vector representations are moved radially to the hovered element in an animation, since showing correct distances the original position of moved items is depicted in light gray to help users preserve their mental map of the initial representation.

Input

- As input, the tool requires a text file with the high-dimensional objects and features, relying on three comma-separated columns per line:
  - ⟨word, feature, co-occurrence frequency⟩
- Optionally, the user may provide text files with the gold standard and/or automatic class assignments, relying on two comma-separated columns per line: ⟨word, class⟩.