A Corpus-Based Study on the Syntactic Behaviour of German Particle Verbs

Nana Khvtisavrishvili, Stefan Bott & Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung

Introduction

Particle Verbs (PVs) consist of two parts: a particle and a base verb (BV). They can occur syntactically separated, morphologically separated or written together.

Preferances of different PVs for different syntactic paradigms. There is a research interest in finding out which factors affect the frequency areas - high, mid and low. 90 PVs for each particle, except unter(38 PVs). Occurrence frequencies are calculated as the normalized harmonic mean of four different frequencies gained from the following corpora: SdeWaC, HGC, COW and Wikipedia.

Materials and Tool

SdeWaC [3]: collection of German texts from German web pages; ca. 880 million tokens; parsed with Bohnets MATE dependency parser [1]. Used to gain occurrence frequencies of PV syntactic paradigm.

Gold standards: some for each hypothesis

Dataset: 938/629 PVs, selected randomly from three frequency areas - high, mid and low. 90 PVs for each particle, except unter(38 PVs). Occurrence frequencies are calculated as the normalized harmonic mean of four different frequencies gained from the following corpora: SdeWaC, HGC, COW and Wikipedia.

Method

Clustering, unsupervised machine learning approach is used. It groups elements with similar feature values in the same cluster.

We use simple K-means: a flat, hard, exhaustive clustering algorithm. It uses squared error criterion:

\[ E = \sum_{j=1}^{n} \sum_{j=1}^{m} |x_{ij} - c_{j}|^2 \]  

The goal of the algorithm is to minimize the average squared euclidean distance between the vectors and the centroid of the cluster. The centroid of a cluster is defined as the mean of the vectors within a cluster:

\[ c_{j}(w) = \frac{1}{|w|} \sum_{i \in w} x_{ij} \]  

PVs are represented in terms of 6-dimensional feature vectors. The normalized frequencies of PV per syntactic paradigm are taken as classification features.

Materials and Tool

SdeWaC [3]: collection of German texts from German web pages; ca. 880 million tokens; parsed with Bohnets MATE dependency parser [1]. Used to gain occurrence frequencies of PV syntactic paradigm.

Gold standards: ones for each hypothesis

Dataset: 938/629 PVs, selected randomly from three frequency areas - high, mid and low. 90 PVs for each particle, except unter(38 PVs). Occurrence frequencies are calculated as the normalized harmonic mean of four different frequencies gained from the following corpora: SdeWaC, HGC, COW and Wikipedia.

Weka [4]: data mining software.

Error Analysis

Parser errors lead to false frequency information

• False POS tags: e.g. TKVZ (separable verbal particle) instead of APPR (preposition) or “durchrauschen” tagged as VVFIN instead of VVIZU

• Inflected forms instead of lemmas: e.g. “aufzumachen” as a lemma of VVIZU form “aufzumachen”

• Incorrect lemmas: e.g. “aufgeschnitten” as a lemma of “aufgeschnitten”

• Ambiguous lemmas: e.g. “zugestehen/zustehen”

Results

Although the hypotheses could not be proven, the results of all experiments are slightly better than the random clustering. The results of 11 particle experiments are better for the hypothesis-while the results of 7 particle experiments are better for hypotheses about PV frequency and ambiguity.

References


Hypotheses

• Particle: an, auf, aus, nach, ab, zu, ein, über, unter, um, durch

• Ambiguity: - H, M, L

• Frequency: - H, M, L

• Synonyms of PVs

• Register

• Baseverb

Evaluation

• Purity(PI): metric based on majority class principle

• Rand Index (RI): pairwise comparison of elements

• Adjusted Rand Index (ARI): RI corrected for chance