Semantische Beziehungen in Texten mit Word2Vec

und der Vergleich zwischen allgemeinen und domänenspezifischen Korpora als Trainingsdaten

Agenda

Motivation und Problemstellung

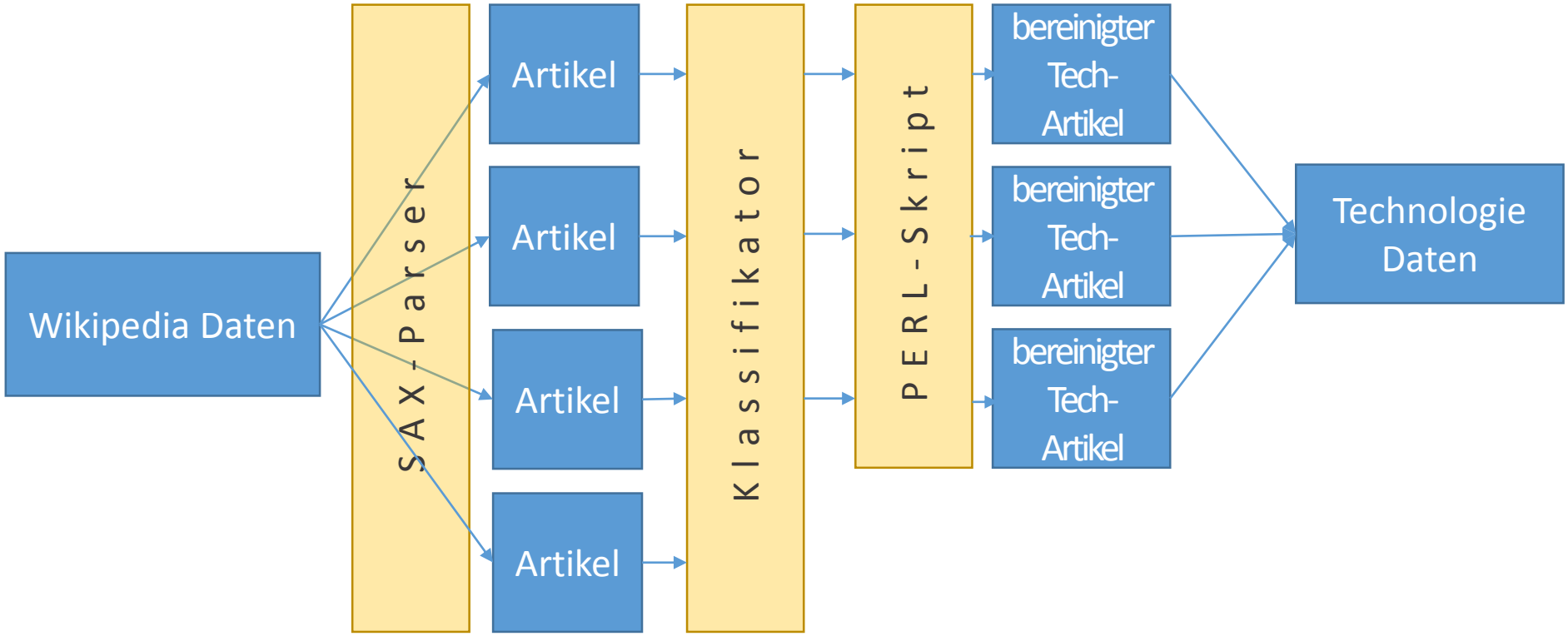
- Vergleich unterschiedlicher Korpora als Trainingsdaten
- Ziel: passende Trainingsdaten für jeweilige Anwendung finden/benützen

Word2Vec

Daten und Vorverarbeitung

- Trainingsdaten: > Milliarden Wörter
- Gesamte Wikipedia Daten
- Teilkorpus über Technologie/PC/Internet

Daten und Vorverarbeitung



Wikipedia-Korpus

- Große Menge an Daten (2,9 Mrd. Wörter)
- Gute Qualität der Artikel
- Breite Menge an Themen
- Semantische Ähnlichkeit: Skip-gram >> CBOW
- Hierarchical softmax: gut für seltene Wörter
- Negative sampling: niedrigdimensionale Vektoren

Wikipedia-Korpus

Size	Window	Min_count	Gesamtaccuracy
400	10	5	53,5% (7027/13144)
400	10	10	52,5% (6905/13144)
300	10	5	52,5% (6860/13144)
300	10	10	52,9% (6951/13144)
200	10	5	50,5% (6632/13144)
200	10	10	50,5% (6636/13144)
100	10	5	42,0% (5517/13144)
100	10	10	41,3% (5431/13144)

Experimente

Fazit und Ausblick

Vielen Dank für die Aufmerksamkeit