
TECHNISCHE HOCHSCHULE DEGGENDORF
THD
Financial Institution
River Bank
„Bank“
Storage
Slope
SEMIARID
MITTELS WORD-SENSE INDUCTION (WSI) WERDEN DIE BEDEUTUNGEN VON BELIEBIGEN WÖRTERN AUS NATÜRLICHSPRACHLICHEN DOKUMENTEN IDENTIFIZIERT UND SOMIT SEMANTISCH DURCHSUCHBAR.
SEMIARID
Weitere Infos

Abstract

Projekttitel/ Project title:

SEMIARID - Natürlichsprachliche Semantische Suche in Big Data

Kurztitel/ Short title:

SEMIARID

Einleitung/ Introduction:

Ziel des Projekts ist die Erforschung, Entwicklung, und Integration von KI-Methoden zur Realisierung einer natürlichsprachlichen semantischen Suchmaschine für Anwendungen in Unternehmen. Speziell wird untersucht, wie Mitarbeitern der Zugang internen Konzerninhalte und Fachinformation durch Suchmaschinen verbessert werden kann.

Ziel/ Aim:

Im Ergebnis soll die Qualität der Suche für das gesamte Spektrum von Benutzer-Anfragen von einer klassischen Stichwortsuche bis hin zu natürlichsprachlichen Fragen und kompletten Problembeschreibungen deutlich verbessert werden.

Methode/ Method:

Zur semantischen Analyse der Suchanfragen ist unter anderem die Bestimmung der Bedeutung der Keywords im Kontext der konkreten Suchanfrage mittels *Word-Sense Disambiguation* (WSD) notwendig. Hierzu werden mittels *Word-Sense Induction* (WSI) zunächst alle Bedeutungen relevanter Wörter aus natürlichsprachlichen, konzerninternen Dokumenten zuverlässig identifiziert und somit die automatische Erstellung von Thesauren für fachspezifische sowie konzerninterne Begriffe ermöglicht.

Ergebnis/ Result:

Durch die Anwendung der von Transformer-Modellen gewonnen Erkenntnisse auf die vorangehende Word2Vec Architektur, konnten bisherige unüberwachte Word2Vec Ansätze verbessert werden. Gleichzeitig bleiben geringerer Ressourcenverbrauch und Nachvollziehbarkeit der älteren Word2Vec Architektur ein deutlicher Vorteil, während die Ergebnisqualität der Word2Vec-Ansätze die der Transformer-Modelle approximiert. Zusätzlich wurde ein Parameter-Score für das Clustering der Bedeutungen mittels DBSCAN vorgestellt, anhand dessen die Parameterauswahl des Clustering-Verfahrens vollständig automatisiert wurde.

Projektbeteiligte/ Project participants:

TH Deggendorf – Prof. Dr. Andreas Fischer
TH Deggendorf – Johannes Reisinger

TH Deggendorf – Zineddine Bettouche

Projektpartner/ Project partners:

IntraFind Software AG
DATEV eG

Gefördert durch/ Funded by:

Bayerisches Verbundforschungsprogramm (BayVFP)
des Freistaates Bayern Förderlinie "Digitalisierung"
VDI / VDE / IT

Logos/ Logos:

insgesamt maximal 450 Wörter/ limit of 450 words in total