Abstract

Eine reichhaltige Auszeichnung mit Metadaten ist für alle Arten von Korpora für die linguistische Forschung wünschenswert. Für große Korpora (insbesondere Webkorpora) müssen Metadaten automatisch erzeugt werden, wobei die Genauigkeit der Auszeichnung besonders kritisch ist. Wir stellen einen Ansatz zur automatischen Klassifikation nach Themengebiet (*Topikdomäne*) vor, die auf dem lexikalischen Material in Texten basiert. Dazu überführen wir weniger gut interpretierbare Ergebnisse aus einer sogenanntem *Topikmodellierung* mittels eines überwachten Lernverfahrens in eine besser interpretierbare Kategorisierung nach 13 Themengebieten. Wir evaluieren das Verfahren auf Webtexten aus DECOW14 und Zeitungstexten aus DeReKo, für die jeweils getrennte Goldstandard-Datensätze manuell annotiert wurden.