



schlichen Bewertungen korrelieren als andere Maße, die auf WordNet und ähnlichen semantischen Ressourcen beruhen. Anschließend empirische Arbeiten zu Themenkohärenzmaßen [6,9,5] schlugen eine Vielzahl von weiteren Maßen vor, die auf Wortstatistiken basieren. Diese Maße unterscheiden sich in mehreren Details wie der Definition, Normalisierung und Zusammenfassung von Wortstatistiken sowie den Referenzkorpora zur Erstellung der Statistiken. Weiterhin wurde kürzlich in [1] eine neue Methode basierend auf Kontextvektoren vorgeschlagen. Die Beiträge zur Kohärenz aus Wissenschaftsphilosophie und Text Mining sind komplementär. Während in wissenschaftsphilosophischen Beiträgen Schemata zum Vergleichen von Fakten vorgeschlagen werden, entwickeln die Text-Mining-Beiträge Methoden zum Schätzen und Zusammenfassen von Wortwahrscheinlichkeiten. Es fehlt jedoch eine systematische Evaluation der Methoden aus beiden Bereichen und deren noch unerforschten Kombinationen.

Menschliche Themen-Rankings dienen als Goldstandard für die Evaluation von Kohärenz, die jedoch aufwendig zu erstellen sind. Unsere empirische Evaluation nutzt alle drei öffentlich verfügbaren Quellen solcher Rankings: (1) die Daten von Chang et al. [3], die von Lau et al. [5] für Kohärenzevaluation vorbereitet wurden, (2) Aletras und Stevenson [1] und (3) Rosner et al. [8]. Die Beiträge dieser Arbeit sind: erstens, wir schlagen einen vereinheitlichenden Rahmen vor, der einen Konfigurationsraum aufspannt, der alle bekannten Kohärenzmaße und die Kombinationen der einzelnen Ideen der Ansätze enthält. Zweitens, der Konfigurationsraum wird systematisch durchsucht und alle Kohärenzmaße, bekannte und bisher nicht bekannte, werden auf den verfügbaren Benchmark-Daten evaluiert. Die Ergebnisse zeigen, dass eine bisher nicht bekannte Kombination von Ideen bisheriger Ansätze deutlich stärker mit menschlichen Bewertungen korreliert. Abschließend diskutieren wir Anwendungen von Kohärenzmaßen, die über Themenmodelle hinausgehen.

## Literatur

1. Aletras, N., Stevenson, M.: Evaluating topic coherence using distributional semantics. In: Proc. of the 10th Int. Conf. on Computational Semantics (IWCS'13). pp. 13–22. (2013)
2. Bovens, L., Hartmann, S.: Bayesian Epistemology. Oxford University Press (2003)
3. Chang, J., Gerrish, S., Wang, C., Boyd-graber, J.L., Blei, D.M.: Reading tea leaves: How humans interpret topic models. In: Advances in Neural Information Processing Systems 22, pp. 288–296. (2009)
4. Douven, I., Meijs, W.: Measuring coherence. *Synthese* 156(3), 405–425 (2007)
5. Lau, J.H., Newman, D., Baldwin, T.: Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: Proc. of the European Chapter of the Association for Computational Linguistics (2014)
6. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proc. of the Conf. on Empirical Methods in Natural Language Processing. pp. 262–272. (2011)
7. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conf. of the North American Chapter of the Association for Computational Linguistics. pp. 100–108. (2010)
8. Rosner, F., Hinneburg, A., Röder, M., Nettle, M., Both, A.: Evaluating topic coherence measures. CoRR abs/1403.6397 (2014), <http://arxiv.org/abs/1403.6397>
9. Stevens, K., Kegelmeyer, P., Andrzejewski, D., Buttler, D.: Exploring topic coherence over many models and many topics. In: Proc. of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. pp. 952–961. (2012)