Evaluation des Konfigurationsraumes von Kohärenzmaßen für Themenmodelle*

Extended Abstract

Michael Röder[1,3], Andreas Both[3] und Alexander Hinneburg[2]

[1] Universität Leipzig
[2] Martin-Luther-Universität Halle-Wittenberg
[3] R&D, Unister GmbH, Leipzig
{michael.roeder|andreas.both}@unister.de,
hinneburg@informatik.uni-halle.de

Eine Menge von Aussagen oder Fakten wird als kohärent angesehen, wenn sie sich gegenseitig unterstützen. Deshalb kann eine kohärente Faktenmenge gut in einem Kontext interpretiert werden, der alle oder die meisten Fakten umfasst. Ein Beispiel für eine solche Faktenmenge ist "das Spiel ist eine Mannschaftssportart", "das Spiel wird mit einem Ball gespielt", "das Spiel erfordert große physische Anstrengungen", die z.B. im Kontext von Fußball einen Sinn ergibt. Eine offene Forschungsfrage ist, wie die Kohärenz einer Faktenmenge quantifiziert werden kann [2]. In Arbeiten aus dem Bereich der Wissenschaftsphilosophie wurden Maße vorgeschlagen, die als Funktionen von Verbund- und Randwahrscheinlichkeiten formalisiert wurden, welche den Fakten zugeordnet sind. Bovens und Hartmann [2] diskutieren viele Beispiele, die zu einer Menge von notwendigen Bedingungen führen, die ein solches Maß erfüllen soll. Die Arbeiten in diesem Bereich beschäftigen sich vor allem mit verschiedenen Schemata, die das Zusammenhängen und zueinander Passen der einzelnen Fakten einer größeren Faktenmenge abschätzen. Beispiele für solche Schemata sind (1) vergleiche jeden einzelnen Fakt mit dem Rest aller verbleibenden Fakten, (2) vergleiche alle Paare von Fakten miteinander und (3) vergleiche alle disjunkten Teilmengen der Fakten miteinander. Diese theoretischen Arbeiten aus dem Bereich der Wissenschaftsphilosophie – siehe [4] für einen Überblick – sind in der Informatik weitgehend unbekannt.

Das Interesse an Kohärenzmaßen entstand im Bereich Text Mining, weil unüberwachte Lernmethoden, wie z.B. Themenmodelle, keine Garantie dafür geben, dass ihre Ausgabe interpretierbar ist. Themenmodelle lernen unüberwacht Themen, die üblicherweise als Menge von wichtigen Wörtern repräsentiert werden. Dies ist eine attraktive Methode, um unstrukturierte Textdaten mit einer Struktur zu versehen. In der grundlegenden Arbeit von Newman et al. [7] werden Kohärenzmaße vorgeschlagen, die bewerten, wie verständlich durch Wortmengen repräsentierte Themen sind. Die vorgeschlagenen Maße behandeln Wörter als Fakten und nutzen das Schema, das paarweise alle Wörter vergleicht. Für die Evaluationen in [7] werden durch Menschen erstellte Themen-Rankings verwendet. Die Auswertungen zeigten, dass Maße, die auf Statistiken über das gemeinsame Auftreten von Wörtern beruhen, stärker mit menschlichen Bewertungen korrelieren als andere Maße, die auf WordNet und ähnlichen semantischen Ressourcen beruhen. Anschließende empirische Arbeiten zu Themenkohärenz-

maßen [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 Wissenschaftsphilosopie 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 und empirische 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 Beträge 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., Nettling, 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)