Joao Quintanilha
11/30/24
CS 119 - Big Data

# Reading 07: Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence

## Category:

This paper falls under the category of an analysis of an existing system, as it critically examines the evaluation frameworks for topic models, particularly the validity of automated coherence metrics in comparison to human judgment tasks. The authors perform a meta-analysis of contemporary neural topic modeling literature and investigate gaps in standardization and validation in existing evaluation practices. While the paper involves measurements to assess models and metrics, these measurements are secondary to its primary goal of analyzing and questioning the reliability of established systems and methods in topic model evaluation. The described methods, such as NPMI and human judgment tasks, were already known and widely used at the time of publication, positioning this paper as a systematic critique and analysis rather than an introduction of new prototypes or primarily a measurement study.

---

## Context:

The paper builds on prior work in both classical and neural topic modeling evaluation. Key related papers include Chang et al. (2009), which introduced the word intrusion task to assess topic coherence, and Lau et al. (2014), which established the use of normalized pointwise mutual information (NPMI) as an automated metric correlated with human evaluations. The authors also reference work by Mimno et al. (2011), which optimized semantic coherence in topic models and highlighted the divergence between perplexity and interpretability. More recent studies, such as Doogan and Buntine (2021), critique existing coherence measures for their incompatibility with modern models, providing a direct influence on this paper's call to reassess evaluation paradigms. Additionally, Burkhardt and Kramer (2019) and Dieng et al. (2020) are referenced for their advancements in neural topic modeling approaches like Dirichlet-VAE and Embedded Topic Models, which are included in the authors' empirical analysis. The meta-analysis of evaluation practices draws heavily on comprehensive reviews like Zhao et al. (2021b), which survey neural topic models and their evaluation trends.

The paper's theoretical bases include the concept of topic coherence, as defined by the ability of topic words to collectively indicate a coherent latent category, and the reliance on automated coherence metrics, particularly NPMI, as proxies for human evaluation. It also leans on the principles of content analysis (Krippendorff, 2004), which frame topic modeling as a tool for summarizing and interpreting large corpora. The study emphasizes the importance of alignment between human interpretability and machine-derived metrics, critiquing the generalization of NPMI beyond its validation with classical models. Another critical theoretical base is Goodhart's Law, which posits that when a measure becomes a target, it loses validity as a measure—used here to challenge the over-reliance on automated metrics. The authors also engage with ideas from statistical power analysis (Card et al., 2020) to ensure robustness in their human evaluation methodology. These foundational concepts underpin their critique of evaluation gaps and inform their argument for more standardized and human-centric evaluation practices in topic modeling research.

---

## Concepts:

**Direct ratings:** involve human evaluators scoring the quality of a topic by reviewing the top words associated with that topic. Typically, evaluators assign a score based on an ordinal scale, such as a three-point system, to rate how coherent or meaningful the topic appears. The approach assumes that the interpretability of a topic can be judged by how well the listed words collectively represent a single, identifiable category. For example, evaluators might score a topic containing the words “dog,” “cat,” “pet,” and “animal” as highly coherent. Direct ratings are a straightforward method to gauge human perception of topic quality and have been used in foundational studies like Newman et al. (2010) and Aletras and Stevenson (2013).

**Intrusion:** first introduced by Chang et al. (2009), is a behavioral task designed to assess topic coherence by challenging evaluators to identify an “intruder” word that does not belong in a set of top-ranked words for a given topic. For example, if a topic contains “apple,” “banana,” “fruit,” “orange,” and “car,” the word “car” would be the intruder. This task assumes that a coherent topic will make the intruder word easier to spot. The percentage of correct intruder identifications is then used as a measure of the topic’s coherence. The intrusion method is particularly valuable because it tests the strength of word relationships within a topic in a practical, interpretable way, aligning with the human ability to recognize categorical inconsistencies.

**Coherence:** refers to the degree to which the top words in a topic collectively form a meaningful and interpretable group to human evaluators. It represents how well the words "stick together" in terms of indicating a recognizable latent category or concept. For example, a coherent topic might include words such as "dog," "cat," "pet," and "animal," as these words are strongly related and contribute to a single identifiable theme. Coherence is critical in topic modeling because it determines the utility of the model for tasks like content analysis or summarizing large corpora. Human evaluations of coherence often involve tasks such as direct ratings or word intrusion to assess how well a topic aligns with human interpretations of relatedness.

**NPMI (Normalized Pointwise Mutual Information):** is an automated metric commonly used to evaluate topic coherence by measuring the statistical association between pairs of words in a topic. It calculates how often two words co-occur in a given corpus compared to their individual occurrences, normalized to a range between -1 and 1. Higher NPMI values indicate stronger word associations, with 1 representing perfect co-occurrence and 0 indicating no association. The NPMI for a topic is computed as the average NPMI across all pairs of top words within that topic. This metric assumes that higher word association corresponds to greater topic coherence. While it has been widely adopted due to its correlation with human judgments in classical topic models, this paper critiques its validity for neural topic models, noting that it may favor esoteric or corpus-specific topics that are not necessarily interpretable to humans.

---

## Conclusion:

The paper is generally well-written and structured, presenting a complex topic in a way that is accessible to readers with a background in machine learning and natural language processing. The authors clearly define key terms and concepts, such as “Skip-gram model,” “Negative Sampling,” and “subsampling,” which are critical to understanding the approach. Additionally, the paper effectively uses graphs and tables to illustrate empirical results and model performance, particularly in showing the effects of different sampling techniques on training speed and accuracy. The analogies and examples are helpful in conveying how vector operations capture semantic relationships. However, certain sections, such as the mathematical notation in Negative Sampling and hierarchical softmax, could have benefited from additional explanatory details or simplified notation to enhance readability. Overall, the paper is clear, but it requires some prior knowledge, which might limit accessibility for those new to the field.