Session 5. A Comparative Study of Three Strategies

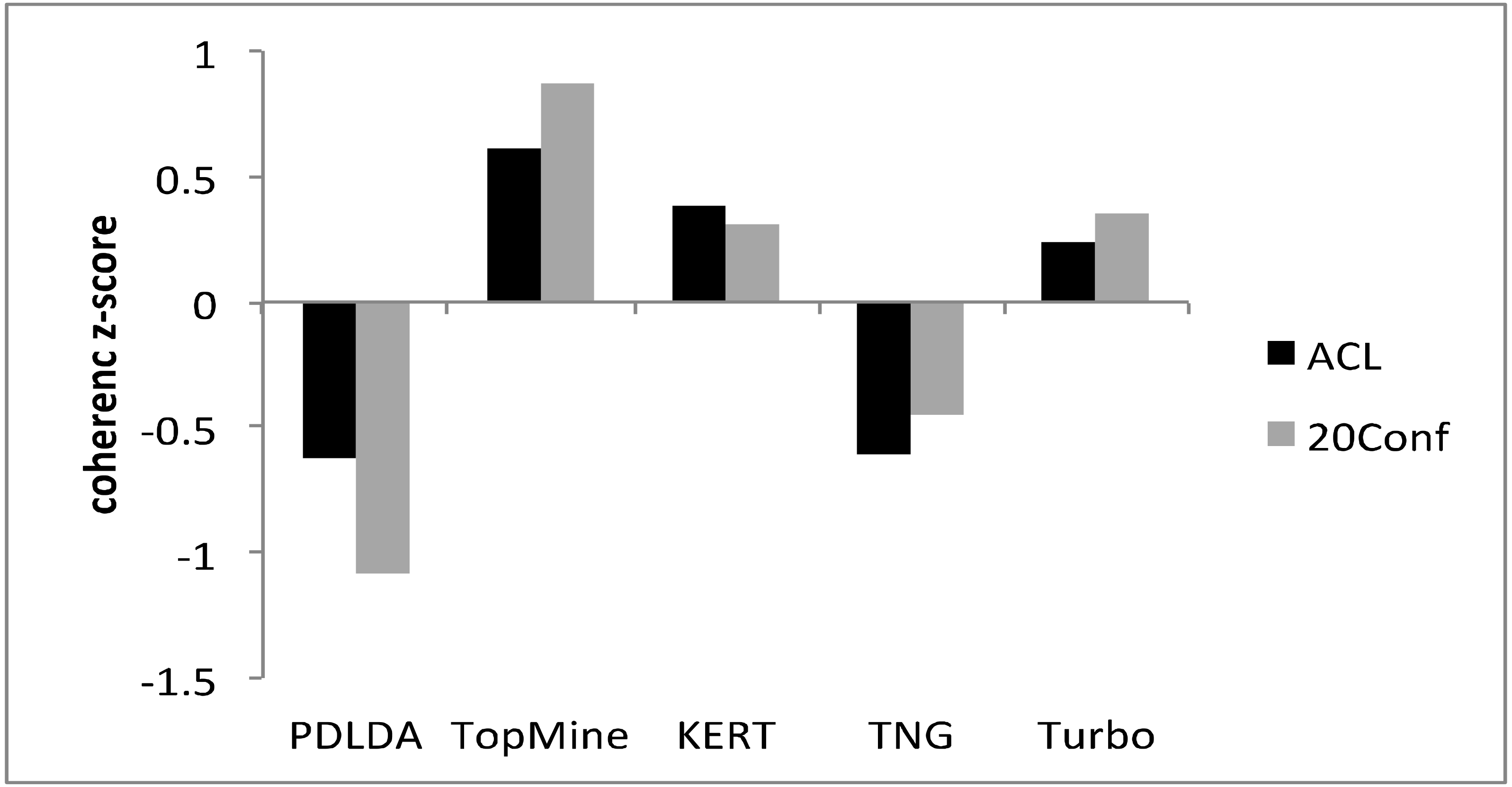
Efficiency: Running Time of Different Strategies

Method	*sampled dblp titles (k=5)*	*dblp titles (k=30)*	*sampled dblp abstracts*	*dblp abstracts*
PDLDA	3.72(hrs)	~20.44(days)	1.12(days)	~95.9(days)
Turbo Topics	6.68(hrs)	>30(days)*	>10(days)*	>50(days)*
TNG	146(s)	5.57 (hrs)	853(s)	NA†
LDA	**65(s)**	3.04 (hrs)	353(s)	13.84(hours)
KERT	68(s)	3.08(hrs)	1215(s)	NA†
ToP-Mine	67(s)	**2.45(hrs)**	**340(s)**	**10.88(hrs)**

Running time: strategy 3 > strategy 2 > strategy 1 (">" means outperforms)

- Strategy 1: Generate bag-of-words → generate sequence of tokens
- Strategy 2: Post bag-of-words model inference, visualize topics with n-grams
- Strategy 3: Prior bag-of-words model inference, mine phrases and impose to the bag-of-words model

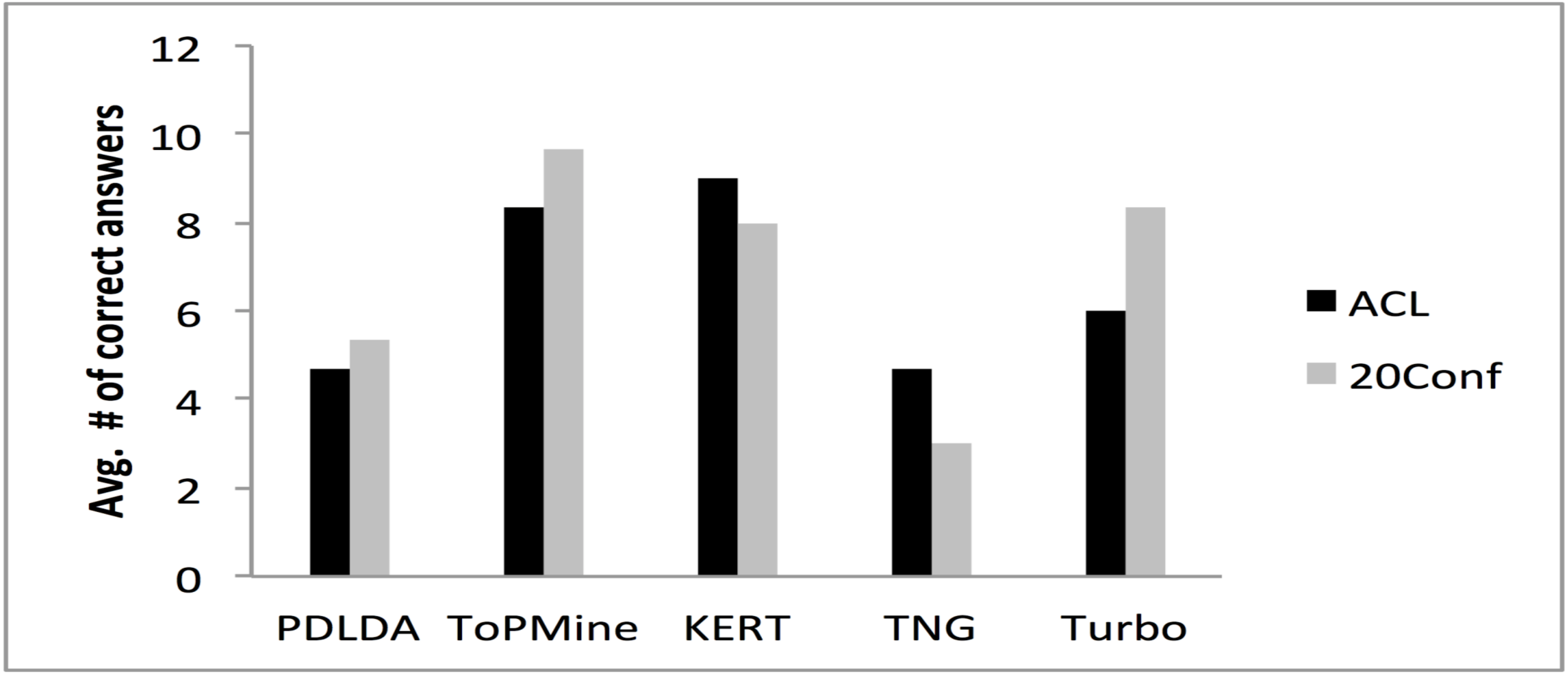
Coherence of Topics: Comparison of Strategies



Coherence measured by z-score: strategy 3 > strategy 2 > strategy 1

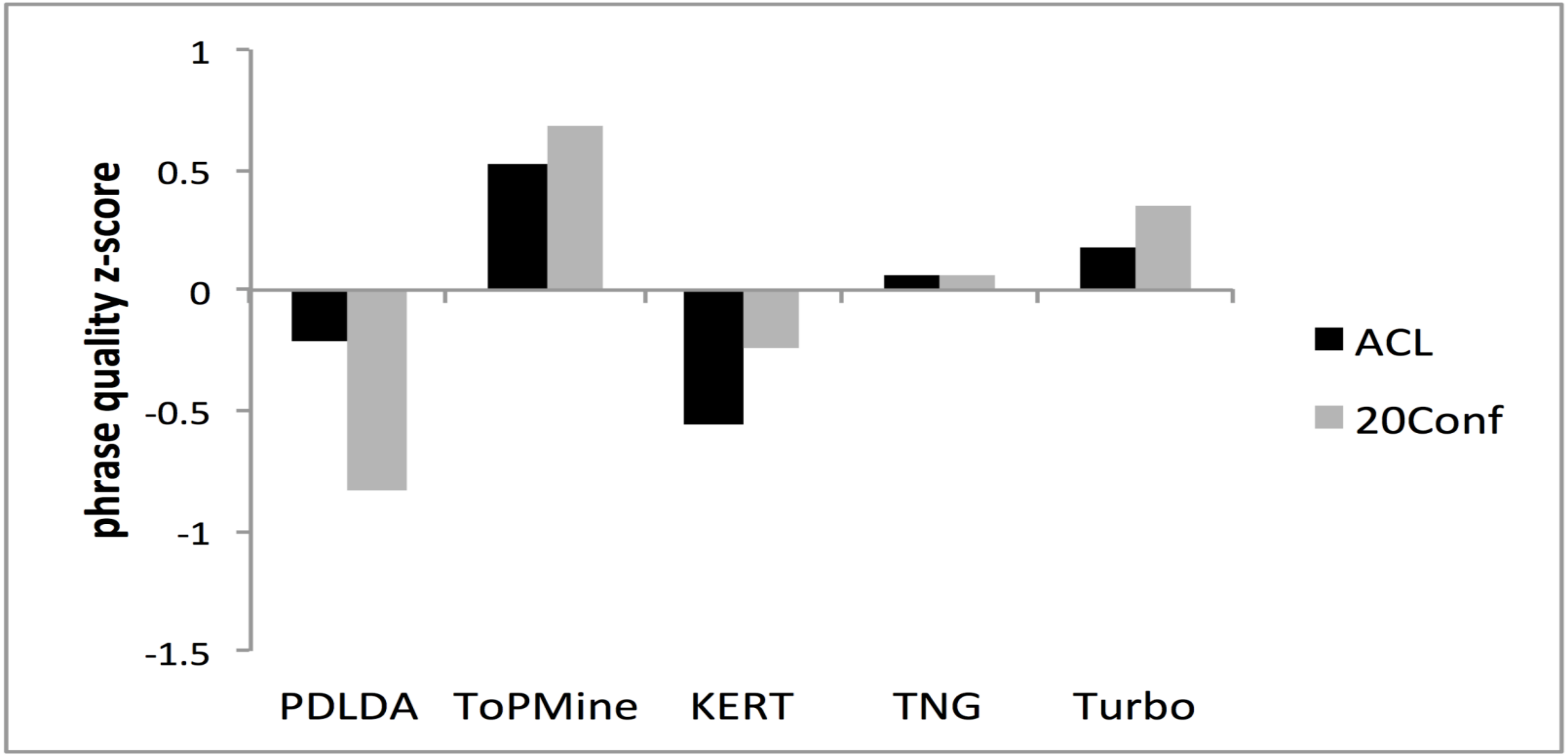
- Strategy 1: Generate bag-of-words → generate sequence of tokens
- Strategy 2: Post bag-of-words model inference, visualize topics with n-grams
- Strategy 3: Prior bag-of-words model inference, mine phrases and impose to the bag-of-words model

Phrase Intrusion: Comparison of Strategies



Phrase intrusion measured by average number of correct answers: strategy 3 > strategy 2 > strategy 1

Phrase Quality: Comparison of Strategies



Phrase quality measured by z-score:
strategy 3 > strategy 2 > strategy 1

Summary: Strategies on Topical Phrase Mining

- Strategy 1: Generate bag-of-words → generate sequence of tokens
  - Integrated complex model; phrase quality and topic inference rely on each other
  - Slow and overfitting
- Strategy 2: Post bag-of-words model inference, visualize topics with n-grams
  - Phrase quality relies on topic labels for unigrams
  - Can be fast; generally high-quality topics and phrases
- Strategy 3: Prior bag-of-words model inference, mine phrases and impose to the bag-of-words model
  - Topic inference relies on correct segmentation of documents, but not sensitive
  - Can be fast; generally high-quality topics and phrases

Recommended Readings

- M. Danilevsky, C. Wang, N. Desai, X. Ren, J. Guo, J. Han. Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents“, SDM’14
- X. Wang, A. McCallum, X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval, ICDM’07
- R. V. Lindsey, W. P. Headden, III, M. J. Stipicevic. A phrase-discovering topic model using hierarchical pitman-yor processes, EMNLP-CoNLL’12.
- Q. Mei, X. Shen, C. Zhai. Automatic labeling of multinomial topic models, KDD’07
- D. M. Blei and J. D. Lafferty. Visualizing Topics with Multi-Word Expressions, arXiv:0907.1013, 2009
- M. Danilevsky, C. Wang, N. Desai, J. Guo, J. Han. Automatic Construction and Ranking of Topical Keyphrases on Collections of Short Documents, SDM’14
- A. El-Kishky, Y. Song, C. Wang, C. R. Voss, J. Han. Scalable Topical Phrase Mining from Text Corpora, VLDB'15
- K. Church, W. Gale, P. Hanks, D. Hindle. Using Statistics in Lexical Analysis. In U. Zernik (ed.), Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon. Lawrence Erlbaum, 1991